HARDNESS IS NOT ENOUGH: WHAT CONTRASTIVE ERROR AUDITS IDENTIFY

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive error auditors rank the items of a data pool by their resemblance to synthetic corruptions, and harder, more plausible corruptions are generally preferred. We prove that no criterion computed from population testing difficulty can certify how well such an auditor ranks real errors. With the pool, error prevalence, representation and real error process held fixed, two corruptions can share the complete binary testing region, and hence the optimal ROC curve, every binary Bayes risk and every f-divergence, while their optimal scores rank the real errors optimally and in exact reverse. This holds on every atomless pool with a nonconstant error posterior, at every continuous difficulty level, and an exact criterion decides it on finite pools. Fidelity, in contrast, is sufficient: audit regret is bounded by the total-variation distance between the corruption and the error distribution plus the contrastive AUC shortfall. Crossed detector interventions on knowledge graphs, including a prespecified replication in an independent architecture, show that the preferred training corruption depends on the errors evaluated, and changing only those errors reverses the order of two published detectors on FB15k-237, by 0.008 in top-5% precision. The dependence persists in image audits at matched measured difficulty and in human-annotated summaries. Synthetic corruptions should therefore be justified against the target error process and evaluated on held-out target errors alongside controls fitted on the same labels.

1 INTRODUCTION

An error auditor ranks the items of a data pool, such as claims extracted into a knowledge graph or generated summaries, so that erroneous items appear first. Because labelled errors are scarce, a common approach trains a classifier to distinguish the pool from synthetic corruptions of it and ranks items by their resemblance to the corruptions. Replacing an entity in a knowledge-graph claim at random, or with an entity of the same type, yields different training tasks and, when used for evaluation, different benchmarks (Zhang et al., 2022; Liu et al., 2024). Harder, more plausible corruptions are generally preferred, as contrastive training negatives (Robinson et al., 2021; Kalantidis et al., 2020) and as the plausible false items that define anomaly and evaluation benchmarks (Liu et al., 2024; Safavi & Koutra, 2020; Zellers et al., 2018). Summary-factuality auditors are likewise trained on rule-based corruptions (Kryscinski et al., 2020), although such synthetic errors can differ from the errors systems make (Goyal & Durrett, 2021).

We ask whether this difficulty, which we term *contrastive hardness*, determines how well the resulting auditor ranks real errors, and prove that it does not. Theorem 1 fixes the pool, the error prevalence, the representation and the latent error process (the unknown distribution of the real errors), and constructs two corruptions with identical complete binary-testing regions: every achievable pair of false- and true-positive rates coincides. Their optimal contrastive scores nevertheless rank the errors in opposite orders, one optimally and one in reverse, so no criterion computed from their testing regions, with any amount of data or training, can determine which corruption to trust. The two corruptions remain different distributions, so criteria that use their semantics, geometry or target labels are not ruled out.

Significance of the complete testing region. That the corruption changes the contrastive optimum is immediate from noise-contrastive estimation. Whether any measure of difficulty can distinguish

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.

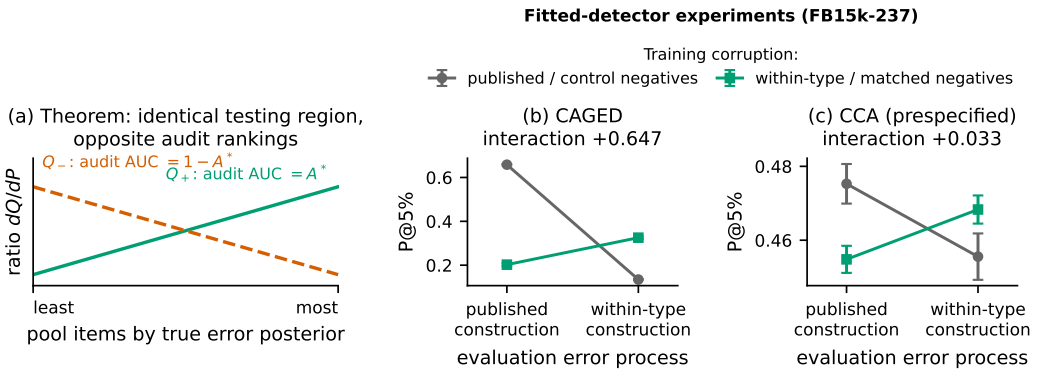

Figure 1: **Identical training difficulty, opposite audit rankings.** (a) Population construction (Theorem 1): the corruption-to-pool ratios $g(U)$ and $g(1-U)$ have the same distribution under the pool, yielding identical complete testing regions but optimal and reversed audit rankings. (b–c) Fitted-detector experiments with CAGED and the prespecified CCA replication on FB15k-237. P@5% is the error fraction among the K highest-scored items, with K equal to 5% of the pre-injection graph for CAGED and of the ranked pool for CCA (their upstream conventions). Points show means and error bars standard deviations over four and five seeds. **Panels (b) and (c) use different vertical scales.** Within-type training changes P@5% by -0.456 and $+0.191$ in (b), and -0.020 and $+0.013$ in (c), on published and within-type errors.

a faithful corruption from an unfaithful one is a separate question. The testing region is the most complete population measure of difficulty: it fixes the optimal ROC curve, every binary Bayes risk and every f-divergence. Two corruptions with opposite audits share it, on any atomless pool and with the errors fixed, so no criterion computed from it, however refined, can establish that an audit exceeds chance on real errors. The construction is unconstrained, so it does not show that realistic corruption families fail. It shows that justifying a corruption by its hardness requires further assumptions about the corruption family or the target, and those assumptions, not hardness, carry the justification. *Fidelity* suffices: a corruption close to the real errors, with a score of small contrastive AUC shortfall, yields small audit regret (Theorem 2).

Contributions. Theorem 1 is the central result, Theorem 2 identifies what replaces hardness, and the experiments show why the distinction matters in fitted audits.

- **Hardness cannot certify audit fidelity (§3.3).** With the pool, prevalence, representation and latent errors fixed, two corruptions with identical complete testing regions have optimal scores attaining audit AUC A^* and $1 - A^*$ (Theorem 1) and the best and worst precision at every inspection budget (Result 1f), on every atomless pool with a nonconstant posterior and at every continuous difficulty level. On finite pools the reversal can fail, and an exact mass-partition criterion decides when.
- **Fidelity suffices and unlabelled data does not (§3.4).** Audit regret is bounded by a corruption-to-error ranking discrepancy, at most their total variation, plus contrastive AUC shortfall (Theorem 2). Unlabelled contrastive data, however plentiful, leaves positive worst-case regret, and a lower bound covers finitely many labels.
- **Corruption choice changes conclusions about published detectors (§4).** In crossed interventions on fitted detectors, including a prespecified replication in an independent architecture, the preferred training corruption depends on the errors evaluated (Figure 1). Changing only those errors reverses the ranking of two published detectors, CAGED and ADKGD, on FB15k-237 by a small margin, though not on WN18RR. The dependence persists at matched measured difficulty in images, and human-annotated summaries show the same interaction without a full crossover. A label-informed correction does not outperform priors fitted on the same labels (Table 1).
- **A reporting protocol for benchmark construction (§5).** We adapt a known maximum-AUC identity into a diagnostic of what a construction lets a cell-only score achieve. Scored

on the detectors' own FB15k-237 benchmark pools and using only the contaminated graph, a learning-free schema-type prior exceeds both detectors' P@5% in every seed, by 0.022 over CAGED and 0.059 over ADKGD on average, so detector results should be reported beside such priors.

2 Related Work

Fixed-target identification, not negative-sampling dependence. Closely related recent work examines objective misalignment in supervised out-of-distribution detection (Li et al., 2025) and risk bounds under auxiliary shift (Wang et al., 2025). Theorem 1 instead concerns identification: it fixes the latent target and the representation and equates the complete testing region. Theorem 2 is a population AUC-regret transfer bound, not a finite-sample risk guarantee. We inherit the NCE optimum (Gutmann & Hyvärinen, 2010), mixture non-identifiability (Bekker & Davis, 2020), density-ratio reweighting (Sugiyama et al., 2007), likelihood-ratio AUC optimality (Clémençon et al., 2008; Menon & Williamson, 2016), binary-experiment equivalences (Reid & Williamson, 2011), the distributional transform (Rüschendorf, 2009) and the categorical maximum AUC (Chen et al., 2021b, Observation 3).

Benchmark artifacts. Partial-input and heuristic baselines expose construction cues in natural-language inference and fact verification (Gururangan et al., 2018; Poliak et al., 2018; McCoy et al., 2019; Schuster et al., 2019), illustrating shortcut learning and benchmark dependence (Geirhos et al., 2020; Dehghani et al., 2021). Adversarial filtering removes easy examples (Zellers et al., 2018; Sakaguchi et al., 2020). Our question is whether the resulting error process matches the audit target. Where a partial-input baseline measures one trained model, the diagnostic of §5 bounds every score that uses only a stated partition of the items.

Contrastive learning and error detection. Negative-sampling studies examine optimization, mining, hard negatives and false negatives (Yang et al., 2020; Wischounig et al., 2026; Robinson et al., 2021; Kalantidis et al., 2020; Chuang et al., 2020). CAGED and CCA (Zhang et al., 2022; Liu et al., 2024), contrastive detectors of erroneous knowledge-graph triples, provide our settings. CCA already motivates semantically plausible anomalies, but it builds benchmark anomalies and training negatives differently (Appendix Q), which motivates our crossed interventions. In summarization, simple synthetic errors like those that train FactCC (Kryscinski et al., 2020) do not reflect the errors of abstractive systems (Goyal & Durrett, 2021), a fidelity gap of the kind Theorem 2 penalizes and Theorem 1 shows hardness cannot detect.

Density models and auxiliary outliers. Auxiliary composition already matters in outlier exposure and its extensions (Hendrycks et al., 2019; Chen et al., 2021a; Choi et al., 2024; Yao et al., 2024), and CRAFT avoids auxiliary outliers (Karunanayake et al., 2025). Perfect density estimation need not detect anomalies (Le Lan & Dinh, 2021), and ranking against a uniform reference on compact support is bipartite ranking (Clémençon & Robbiano, 2014) (Appendix C).

3 What Contrastive Training Identifies

3.1 Contrastive and audit targets

Two rankings. Let p be the density of the observed pool, q that of the synthetic corruption, and p_{err}, p_{pos} those of the errors and the correct items inside the pool. We write P, Q, E and C for the corresponding probability laws, and call Q the *auxiliary*. The population-optimal contrastive anomaly score orders items by q/p (Gutmann & Hyvärinen, 2010), whereas the audit should order them by the likelihood ratio of the error class,

$$\text{contrastive training ranks by } r^{q}_{\mathrm{ctr}}(x) = \frac{q(x)}{p(x)}, \qquad \text{an audit needs } r_{\mathrm{audit}}(x) = \frac{p_{\mathrm{err}}(x)}{p_{\mathrm{pos}}(x)}. \tag{1}$$

The corruption enters the first ratio and not the second.

An item X has latent label $Y \in \{0, 1\}$, with $Y = 1$ denoting error. Let $\pi = \Pr(Y = 1) \in (0, 1)$. Densities below share a reference measure, and we write

$$p = (1 - \pi)p_{\text{pos}} + \pi p_{\text{err}}, \qquad \rho(x) = \frac{p_{\text{err}}(x)}{p_{\text{pos}}(x)}, \qquad \frac{\Pr(Y = 1 \mid x)}{\Pr(Y = 0 \mid x)} = \frac{\pi}{1 - \pi}\rho(x). \tag{2}$$

Where only the correct-class density is zero, take $\rho = +\infty$ and posterior error probability one. Ignore sets with zero pool mass. Thus ρ, the audit ratio r_{audit} of (1), and the posterior $\eta(x) = \Pr(Y = 1 \mid x)$ induce the same ordering. Higher anomaly scores mean more likely errors. For laws F, G write $A(s, F, G) = \Pr(s(X) > s(X')) + \frac{1}{2}\Pr(s(X) = s(X'))$, with $X \sim F$ and $X' \sim G$ independent, for the tie-aware AUC of a score s. The audit AUC is $\text{AUC}(s) = A(s, E, C)$, and A^* is its maximum over measurable scores. Appendix E proves every result, and exact or exhaustive machine checks in the supplement corroborate the proofs. Appendix-only results keep their numbers, such as Result 1e.

Representation and support. All laws refer to the information a score retains: for a feature map ψ, each law is that of $\psi(X)$. Ranking by q/p is defined P-almost surely (zero where $q = 0$), but a finite log-ratio score and density recovery need $q > 0$ P-almost surely, which published detectors can violate: CAGED rejects corruptions that coincide with real graph triples. We therefore do not equate a fitted neural score with a density ratio.

Proposition 1 (contrastive score and matched ordering)**.** *At the population optimum, after removing the training-class-prior constant, a classifier separating pool p from corruption q has logit $\ell_q(x) = \log[p(x)/q(x)]$. Its anomaly score is $d_q = -\ell_q$. If $q = p_{\text{err}}$, then $q/p = \rho/[(1 - \pi) + \pi\rho]$, which is strictly increasing in ρ. More generally, any normalizable $q \propto p\, h(\rho)$ with positive, strictly increasing h gives the oracle ordering.*

3.2 A TWO-ITEM EXAMPLE

Let two item values a, b have pool law $P = (0.5, 0.5)$. Fix prevalence $\pi = 0.5$, error law $E = (0.8, 0.2)$ and correct-item law $C = (0.2, 0.8)$, so $P = \pi E + (1 - \pi)C$ and a is more likely to be an error. Choosing $Q_+ = E$ gives contrastive ratios $Q_+/P = (1.6, 0.4)$, which rank a first and attain audit AUC 0.8. Choosing $Q_- = C$ reverses the scores and gives audit AUC 0.2. Both training tasks have optimal contrastive AUC 0.65. More strongly, swapping a and b preserves P and maps Q_+ to Q_-, mapping every randomized test to one with identical class-conditional acceptance rates. Thus the complete testing regions coincide.

3.3 SAME HARDNESS, OPPOSITE AUDIT RANKINGS

For a pool law P and auxiliary Q, the *testing region* $\mathcal{T}(P, Q)$ contains all achievable false-positive and true-positive rate pairs for telling Q from P. Formally, it is the set of pairs $(\mathbb{E}_P\phi, \mathbb{E}_Q\phi)$ over randomized tests ϕ with values in $[0, 1]$. It determines the optimal ROC curve, every binary Bayes risk and every f-divergence (Reid & Williamson, 2011). It thus captures population difficulty for unrestricted tests, while leaving feature geometry, restricted model classes and optimization outside its scope.

Theorem 1 (same hardness, opposite audit rankings)**.** *Let $P = \pi E + (1 - \pi)C$, where E and C are the error and correct-item laws, be atomless (no individual item has positive probability mass) on a standard Borel space (such as $\mathbb{R}^d$ with its Borel sets), with posterior η not P-almost surely constant, and let g be any strictly increasing probability density on $(0, 1)$. There is a function U of the item, uniform under P and nondecreasing in η, such that $dQ_+/dP = g(U)$ and $dQ_-/dP = g(1 - U)$ satisfy*

$$\mathcal{T}(P, Q_+) = \mathcal{T}(P, Q_-), \qquad \text{AUC}(dQ_+/dP) = A^* > \tfrac{1}{2}, \qquad \text{AUC}(dQ_-/dP) = 1 - A^*. \tag{3}$$

Proof idea. U is the item's quantile position under the posterior η (its distributional transform), with a measurable tie breaker on posterior level sets. Every test on $(P, Q_\pm)$ may be replaced by its conditional expectation given U, and $u \mapsto 1 - u$ maps the tests of one experiment bijectively

onto those of the other, so the regions coincide. U orders pairs with different posteriors as η does, and pairs on one level set with probability one half, matching the tie credit. Reversing a score maps AUC A to $1 - A$ (Appendix E.1).

The density g sets the common difficulty: the closer it is to 1, the harder both corruptions are to distinguish from the pool. Conversely, any pair (P', Q') with $Q' \ll P'$ (Q' puts no mass where P' has none) whose ratio dQ'/dP' has an atomless law has the testing region of some g, so the reversal is available at every continuous difficulty level. Any certificate computed from the testing region, such as optimal contrastive AUC, a divergence or a Bayes risk, therefore returns the same verdict for an optimal and a reversed audit, and any non-abstaining verdict about the sign of audit advantage is wrong for one of them (Appendix E.1). The quantile rearrangement of the posterior is classical (Rüschendorf, 2009). What it yields here is not: one fixed latent target, two corruptions indistinguishable as testing problems, and opposite audit orders, on every atomless pool. Unlike the swap of §3.2, it needs no symmetry, and with indivisible item masses it can fail.

Exact finite-pool characterization. Indivisible pool masses $m_1, \ldots, m_k$ of the k items can prevent reversal. List the distinct posterior levels in increasing order, form their cumulative pool masses S_ℓ ($S_0 = 0$), and cut $[0, 1]$ at every S_ℓ and $1 - S_\ell$. A pair $Q_\pm$ with identical testing regions attaining A^* and $1 - A^*$ exists if and only if the items of each level ℓ can be grouped into sets whose masses equal the lengths of the cut intervals inside $[S_{\ell-1}, S_\ell]$. With distinct posteriors this reduces to $m_i = m_{k+1-i}$ in posterior order, and ties admit further solutions. With ties, deciding feasibility is NP-complete in the weak sense already at two posterior levels, by reduction from PARTITION, and that case has a pseudo-polynomial algorithm (Appendix E.1). Two results extend the reversal beyond these pools. On any pool, Result 1e constructs a pair with equal optimal contrastive AUC, total variation and χ^2 divergence, though not an equal full region. Result 1f shows that such pairs, like the atomless pair, attain the best and worst precision at every inspection budget, the population analogue of P@5%.

Scope of the results. Hard negatives are not claimed harmful: successful contrastive ranking is part of the sufficient condition below. Q_+ and Q_- are different distributions, so methods using semantic structure, feature geometry or target labels escape this obstruction. Theorem 1 is an existence result using the latent posterior. It gives no label-free selection rule and imposes no locality or semantic-edit constraint. Identification within domain-specific corruption families is open.

3.4 WHEN CORRUPTION FIDELITY GUARANTEES ACCURATE AUDITING

Theorem 2 (fidelity bounds audit regret)**.** *For an auxiliary $Q \ll P$ and any score s, let $\epsilon_{\mathrm{ctr}}(s) = \sup_v A(v, Q, P) - A(s, Q, P)$ be its contrastive AUC shortfall, with suprema over measurable scores v. The audit regret $\mathcal{R}(s) = A^* - \mathrm{AUC}(s)$, the AUC lost relative to the optimal error ranking, satisfies*

$$\mathcal{R}(s) \le \min\left\{1, \frac{2d_{\mathcal{G}}(E, Q) + \epsilon_{\mathrm{ctr}}(s)}{1 - \pi}\right\}, \tag{4}$$

where $d_{\mathcal{G}}(E, Q) = \sup_v |A(v, E, P) - A(v, Q, P)| \le \mathrm{TV}(E, Q)$ and $\mathrm{TV}(F, G) = \sup_B |F(B) - G(B)|$ over measurable B.

The first term is target mismatch in the audit's own currency: the largest change in any score's AUC against the pool when the errors are replaced by the corruption. Indicator scores give $d_{\mathcal{G}} \ge \frac{1}{2}\mathrm{TV}(E, Q)$, so total variation loses at most a factor of two. The second term is contrastive ranking error, which better fitting reduces without controlling the first, so training performance alone cannot certify the bound. Q need not be a compatible error component. The bound is informative only when $2d_{\mathcal{G}}(E, Q) + \epsilon_{\mathrm{ctr}}(s) < 1 - \pi$, that is, for auxiliaries already close to the errors.

Proof idea. Since $A(s, E, E) = \frac{1}{2}$, the mixture gives $A(s, E, P) = (1 - \pi)A(s, E, C) + \pi/2$ for every score and for the optima. Replacing E by Q moves this functional by at most $d_{\mathcal{G}}(E, Q)$ for every score, and hence moves its supremum by at most as much. Adding the two changes and the shortfall gives $(1 - \pi)\mathcal{R}(s) \le 2d_{\mathcal{G}}(E, Q) + \epsilon_{\mathrm{ctr}}(s)$.

Results 5e and 5f extend the bound to AUC within observed strata, when the auxiliary matches the pool's stratum masses, and to a change of pool between training and deployment. The pool-shift

Table 1: All studies of §4: upper rows test corruption dependence, lower rows label-informed matching. Effects are P@5% differences for detectors on FB15k-237 and AUC differences otherwise. Each detector effect has one sign in every seed (four for CAGED and ADKGD, five for CCA, with p its prespecified exact sign test). The system margin is ADKGD minus CAGED on published, then within-type, errors. Spans are human-annotated hallucination spans. Intervals are 95%, except QA's Bonferroni 97.5%. The refit row adds fitting uncertainty post hoc. †The protocol's label: the preferred auxiliary reverses on both backbones, but a crossover requires every effect to reach 0.05.

Study	Evaluated errors	Contrast	Effect	Outcome
CAGED	two injected	interaction	$+0.647$	crossover
CCA	two injected	interaction	$+0.033$, $p = 0.031$	crossover
ADKGD/CAGED	two injected	system margin	-0.037, then $+0.008$	order reverses
Images	two natural	matched−other	$+0.044$ to $+0.324$	positive interaction†
FRANK	two human	interaction	$+0.206$ $[0.121, 0.297]$	positive, no crossover
Re-DocRED	human removals	matched−random	$+0.304$ $[0.253, 0.357]$	positive, prior higher
CoNLL++	human revisions	matched−random	$+0.0212$ $[-0.0145, 0.0572]$	unresolved
Summary	human spans	matched−generic	$+0.022$ $[0.005, 0.040]$	positive, not strong
QA	human spans	matched−generic	$+0.048$ $[0.021, 0.077]$	strong, fixed heads
QA (refit)	same sample	same contrast	$+0.048$ $[-0.015, 0.075]$	unresolved
Data2txt	human spans	matched−generic	-0.022 $[-0.041, -0.004]$	transfer fails

term cannot be dropped: an auxiliary equal to the error law, with zero contrastive shortfall on its training pool, still reverses the ranking when that pool differs from the target (Result 5f).

Limits of unlabelled contrastive data. If the target is unknown, the observables (P, Q, π) admit every measurable score ordering through different splits of the pool into errors and correct items (Results 1b–1c). Result 1g quantifies this for two worlds that share pool and prevalence but order the errors oppositely: unlimited unlabelled samples leave worst-case regret $\kappa > 0$. With n annotated labels it is at least $\kappa \max\{0, 1 - \sqrt{nK/2}\}$, for explicit constants $\kappa, K > 0$ (Appendix E.1).

Summary. Hardness alone cannot certify audit fidelity. Fidelity to the target errors, together with successful contrastive ranking, is sufficient for small audit regret.

4 Corruption Choice Changes Audit Conclusions

The experiments ask four questions: does the preferred training corruption depend on the evaluated errors (§4.1), does the dependence persist at matched measured difficulty (§4.2) and extend to human-identified errors (§4.3), and does label-informed matching add value beyond same-label controls (§4.4)? Table 1 summarizes every study, including those that miss their criteria. They test fitted auditors under specified interventions, not the theorem's construction (Appendix C.1). A finite-cell simulation first isolates the mechanism: with the error process fixed and the corruption tilted progressively toward plausible claims, audit AUC peaks where the corruption matches the errors and falls from 0.832 to 0.632 at the most plausible setting (Appendix AA).

A single benchmark confounds a training corruption with the errors that benchmark injects, so the crossed designs pair two training corruptions (columns) with two evaluated error processes (rows). A within-row contrast is the effect of changing corruption. An interaction means that this effect differs between rows, and a crossover additionally requires the preferred corruption to reverse. Detector studies report P@5% and the others AUC. A result is *positive* when its prespecified interval excludes zero in the predicted direction, and *strong* when its effect also reaches a prespecified size.

4.1 Does the preferred corruption depend on the evaluated errors?

In every crossed detector comparison, the preferred training corruption is the one matching the evaluated errors. CAGED uses one generator for injected anomalies and training negatives. We separate the two choices and cross them, holding the rest of the detector fixed. The published generator replaces one uniformly drawn slot with a uniformly drawn filler, and half its anomalies are wholly

random triples. The argument it inserts is thus rarely one the relation is observed with, and a prior on that alone (§5) can reject it. The within-type generator draws the replacement from entities attested in that slot for that relation, reducing this cue while leaving other construction cues possible. At the published configuration the patched code reproduces CAGED's recorded results exactly. On FB15k-237 (Toutanova & Chen, 2015), within-type negatives raise P@5% on the within-type row by 0.191 and lower it on the published row by 0.456 (Figure 1b), with every within-row contrast in the same direction across seeds. The same crossing appears on the knowledge-graph benchmark WN18RR (Dettmers et al., 2018) and in ADKGD, a dual-channel knowledge-graph anomaly detector (Wu et al., 2025) (Appendix Q).

Prespecified replication in an independent architecture. Because CAGED and ADKGD share corruption code, we test the interaction in CCA under a protocol whose inputs, checksums and directional prediction were frozen before any outcome existed. CCA builds its training negatives by swapping entities within a batch, so its matched arm reweights these candidates. A candidate's weight is how often its new entity fills that argument slot of that relation in the public graph, relative to how often it fills that slot overall. A candidate that coincides with a real triple receives zero weight. This moves the negatives toward plausible claims without using any error label. On average, matching raises within-type P@5% by 0.013 and lowers published-error P@5% by 0.020, and each change has that sign in all five seeds. The interaction is 0.033, positive in every seed (Figure 1c), with prespecified one-sided exact sign-test $p = 0.031$, the smallest attainable, conditional on this graph, anomaly draw and pretraining checkpoint.

A reversed system comparison. With published training negatives, changing only the injected error construction reverses the CAGED–ADKGD ordering on FB15k-237. CAGED leads the published benchmark by 0.037 and ADKGD leads on within-type errors by 0.008, each on all four seeds. A detector margin can therefore partly reflect the corruption that defines the benchmark rather than detector quality, although this margin is small and WN18RR does not invert. The pairing is exact because ADKGD forks CAGED's corruption code.

4.2 Does the dependence persist at matched measured difficulty?

It persists, although not at the prespecified strong size. The image study is the closest empirical analogue of the theorem. A contaminated pool (0.90 CIFAR-10, 0.05 each CIFAR-100 natural and SVHN digit images) is audited by linear heads separating it from Tiny ImageNet objects or EMNIST letters, on frozen CIFAR-10-trained features (Krizhevsky, 2009; Netzer et al., 2011; Le & Yang, 2015; Cohen et al., 2017). Each auxiliary Q is diluted with pool images, $Q_\lambda = (1-\lambda)P + \lambda Q$, until the two heads' held-out pool-versus-auxiliary AUCs agree within ± 0.02: $(0.8997, 0.8998)$ for objects and letters with ResNet-18 (He et al., 2016) and $(0.9547, 0.9499)$ with WRN-28-10 (Zagoruyko & Komodakis, 2016), with paired 90% difference intervals $[-0.0048, 0.0047]$ and $[0.0019, 0.0078]$, a measured proxy rather than the population testing region. Because $Q_\lambda/P = (1-\lambda) + \lambda\, Q/P$, dilution changes difficulty but preserves each auxiliary's density-ratio ordering. Semantic fidelity is not isolated from other differences between the sources. At matched difficulty the auxiliary resembling each error type (objects for natural images, letters for digits) leads on all four comparisons. It leads by 0.301 and 0.044 AUC on natural and digit errors for ResNet-18 and by 0.324 and 0.067 for Wide-ResNet-28-10, with every Bonferroni lower bound above zero, so the preferred auxiliary reverses between the two error types on both backbones. The frozen protocol, amended from development data before target data were examined, reserves *crossover* for a reversal in which every effect reaches 0.05. ResNet-18's digit effect misses that size, so the protocol records a positive interaction without a crossover (Appendix Z).

4.3 Does corruption dependence extend to human-identified errors?

The dependence extends to human-identified errors, although asymmetrically. FRANK (Pagnoni et al., 2021) annotates, with its own annotators, entity errors (a misstated primary argument or attribute) and out-of-article errors (information absent from the source) in summaries from nine pre-large-language-model systems. Two label-free auxiliaries mirror that distinction, swapping a name or number for one from the same article or from another, which predicts the preferred auxiliary for each error type. Under a prospective protocol, with every head refitted in each bootstrap replicate,

the interaction is 0.206 $[0.121, 0.297]$. It is carried by out-of-article errors, where the cross-article swap leads by 0.196. The harder within-article swap gains only 0.011 $[-0.026, 0.049]$ on entity errors, so a full crossover is not established (Appendix J).

4.4 Does label-informed matching add value beyond same-label controls?

Not beyond same-label controls. Matching improves on random or generic corruptions in three of four fitted studies, but same-label priors and supervision match or exceed it.

Human revisions of relation annotations. Re-DocRED (Tan et al., 2022), a human revision of DocRED (Yao et al., 2019) made mainly to add missing relations, supplies the errors: the original assertions it removed, with those it retained as correct. One contrastive detector is fixed in every arm and only its training corruption varies, over 10 seeds of document-disjoint halves, with a disclosed post-outcome correction to tie handling and to the bootstrap over documents shared across splits (Appendix X). Uniform within-document replacements give AUC 0.522 and type-preserving replacements 0.436, below chance. The matched arm reweights the same candidates so that their marginal distribution over schema cells (relation, head type, tail type) matches the Jeffreys-smoothed distribution of the removed triples in the fitting half. It reaches 0.826, a gain of 0.304 with the same sign in all 10 seeds. Post-hoc controls qualify this. The smoothing alone, with no labels, gives a gain of 0.148. Permuted labels give 0.136 on average, so the labels add 0.168 beyond permutation, and a relation prior fitted to the same labels reaches 0.879, above matching. On CoNLL++ (Wang et al., 2019b), whose errors are the entity labels changed by a hand-corrected reannotation of CoNLL-2003 (Tjong Kim Sang & De Meulder, 2003), the same prespecified contrast is unresolved (Appendix Y).

Summaries. Summaries from RAGTruth, a human-annotated hallucination benchmark (Niu et al., 2024), are audited under a prospective protocol by identical linear heads on frozen natural-language-inference (NLI) and lexical features, trained against generic mismatches (each summary paired with another source) or against the matched auxiliary (the human-labelled training errors). On 900 official-test responses, matching raises AUC from 0.698 to 0.720, clearing the positive criterion but not the 0.03 strong threshold. Equal hardness was not confirmed, and the controls limit the interpretation: same-label supervision gives 0.720, permuted-label matching 0.717, and a length plus generator/source prior 0.800. An exploratory decomposition attributes most of the gain to cross-generator comparisons, so no semantic-fidelity benefit is isolated (Appendix G).

Question answering and transfer. A prospective QA study gives every auxiliary the pool's generator/source composition and scores AUC only within generator-and-source groups (Result 5e), so composition alone cannot decide the comparison. The matched auxiliary reweights the fitting errors to that composition, the generic one pairs each response with another source's passages, and a control applies the matched weights to labels shuffled within strata. On 2305 responses from 394 held-out source groups, matching reaches 0.732 against 0.684 generic and 0.620 shuffled. Its gains of 0.048 and 0.112 clear the 0.03 strong criterion under the prespecified bootstrap, which holds the fitted heads fixed. Same-label supervision (0.744) and matching without the composition weights (0.761) remain stronger, and a post-hoc bootstrap that also refits the heads leaves the generic contrast unresolved (Appendices H and H.1). Applied unchanged to 6195 Data2txt business descriptions generated from structured records, the matched heads trail generic ones, so the positive criterion fails, which is compatible with Result 5f for a different pool but does not validate it (Appendix I).

5 A Supporting Benchmark Diagnostic

Several priors above rank items by a coarse property alone, such as a claim's relation or entity types. This section bounds what any such score can achieve. A *cell map* T partitions items into k groups, such as relations, type signatures, or bins of lexical features. A cell-only score assigns one value s_a to every item in cell a. Define the class-conditional cell probabilities $e_a = \Pr(T(X) = a \mid Y = 1)$ and $c_a = \Pr(T(X) = a \mid Y = 0)$. These are probabilities of cells within each class, not error rates within cells. For $h_{ab} = \mathbf{1}[s_a > s_b] + \frac{1}{2}\mathbf{1}[s_a = s_b]$, a cell-only score has $\mathrm{AUC}(s) = \sum_{a,b} e_a c_b h_{ab}$.

Proposition 2 (balanced cells remove the shortcut)**.** *If $e = c$, every cell-only score has AUC $\frac{1}{2}$.*

Proposition 3 (maximum cell-only AUC). *Let $D_{ab} = e_a c_b - e_b c_a$ and define*

$$\Lambda(e, c) = \frac{1}{2} \sum_{a<b} |D_{ab}|. \tag{5}$$

Every cell-only score satisfies $|\mathrm{AUC}(s) - \frac{1}{2}| \le \Lambda$. Ordering by e_a/c_a attains $\frac{1}{2} + \Lambda$ (Chen et al., 2021b, Observation 3), as does ordering nonempty cells by $e_a/(e_a + c_a)$, which allows $c_a = 0$.

We use this known identity as a diagnostic on any stated partition, not only types, and contribute the estimation and reporting procedure around it. A finer partition cannot reduce Λ (Result 4a), but a coarse map cannot bound a richer score. If a fraction u of the error mass lies in cells with no correct mass, their indicator attains AUC $\frac{1}{2} + u/2$, so $\Lambda \ge u/2$ (Result 3a), a floor that on sample counts bounds only the empirical Λ.

Estimation and reporting. On the observed counts, $\frac{1}{2} + \widehat{\Lambda}_{\mathrm{plug}}$ is exactly the maximum cell-only AUC, and the plug-in $\widehat{\Lambda}_{\mathrm{plug}}$ can overestimate population separation. The cross-fitted $\widehat{\Lambda}_{\mathrm{cf}}$ fits cell orderings on one half of the data and evaluates them on the other. When every evaluation cell also appears in its fitting half, $\frac{1}{2} + \widehat{\Lambda}_{\mathrm{cf}}$ is the mean held-out AUC of the fitted cell prior, a sensitivity statistic rather than a population ceiling or an empirical maximum (Appendix F). We recommend reporting the map and how it was selected, labelled cell counts, the plug-in maximum with its support floor, and a prior on the same map, defined without the evaluated labels, beside the detector. On NLI and fact-verification benchmarks built to remove known artifacts, the diagnostic's results are mixed, so it is not an externally validated artifact detector, and not every surveyed map was selected before evaluation-label access (Appendices F.1, K and O).

A learning-free prior against published detectors. A schema-type prior types each entity by the Freebase type it most often heads (the first two components of the relation path, such as `/people/person`) and scores a triple by how rarely the graph pairs its relation with its head's and tail's types, excluding the triple's own support. It uses only the contaminated graph and no labels. Scored on the detectors' own FB15k-237 pools, rebuilt with their loader at their four seeds and identical triple for triple, it reaches P@5% 0.680 (AUC 0.943) against CAGED's 0.658 and ADKGD's 0.621, and leads both in every seed, by 0.022 and 0.059 on average (at least 0.018 and 0.041). Types derived from the pre-injection graph, additional benchmark information, raise it to 0.707. On WN18RR, whose relation names carry no schema, CAGED leads an argument-identity prior, 0.495 ± 0.011 against 0.200 (Appendix U).

6 Discussion and Limitations

Implications for published work. CCA motivates semantically plausible anomalies, yet 62.1% of its benchmark's replacement anomalies are type-preserving against 7.7% of its argument-replacement training negatives (Appendix Q). A prior that uses only the contaminated graph outperforms both published detectors on their own FB15k-237 pools (§5), and changing only the injected errors reverses their ordering (§4.1).

Limitations. The theorem does not rest on the experiments, and Table 1 grades each study against its own prespecified criteria. The image and Summary effects miss their strong thresholds, which is not the same as no effect. FRANK's interaction rests on one of its two rows. QA shares its annotation pipeline and generator families with Summary, and its generic contrast is unresolved once the heads are refitted. The CoNLL++ interval spans zero, and unchanged QA heads do not transfer to Data2txt. On Re-DocRED a same-label relation prior is stronger (§4.4), and on a private corpus of claims a language model extracted from social-media posts, an exponential tilt fitted to three annotators' error labels does not outperform a relation rate fitted to the same labels, with non-blind model selection (Appendices S.1 and L).

Conclusion. Identical complete testing difficulty is compatible with the best and worst audit rankings of one fixed error process. Which extra information suffices, whether semantic structure, a small target sample or a corruption rule that outperforms a same-label prior, is open. Contrastive audits should justify the error process they approximate and evaluate correction gains against controls fitted with the same labels.

REPRODUCIBILITY STATEMENT

Complete proofs and their numerical checks are in Appendix E, estimator specifications in Appendix F, and experimental designs, amendments and failed criteria in the corresponding study appendices. Appendix L records model-selection chronology. Reported numerical results are generated from recorded artifacts and checked for agreement. The anonymous diagnostic tool is available at `https://anonymous.4open.science/r/lambda-index-212A`. The supplement contains eight de-identified packages, each with its own verifier: every machine check of the proof map, the per-seed detector results behind Figure 1 and Table 1 with the CCA sign test, intervention source and the prior's paired comparison on the detectors' pools, the protocol and model checksums, and calculation audits and frozen-input replays of Summary, QA, Data2txt and FRANK, including every refitted QA head and optional refitting of all FRANK bootstrap heads. They do not reproduce every upstream training run or the private annotations. Source code for all experiments will be released upon publication. Appendix B details these boundaries.

ETHICS STATEMENT

Annotation. Three annotators judged every row of all three gold sheets, the annotated claims from our private extraction corpus (Appendix S.1). One is an author of this paper and two are colleagues of the authors. All three participated voluntarily and were not compensated. Annotation was performed manually under a written protocol committed before annotation. The protocol permitted one language-model check of each annotator's completed sheet to flag rows for re-reading. The model supplied no labels or proposed verdicts. Annotators reconsidered flagged rows only by rereading the evidence and retained responsibility for every verdict, channel and note. The protocol also let annotators translate evidence sentences in languages they did not read into English with a general-purpose LLM chatbot. Every translated row was marked, and its verdict remained the annotator's own. Disagreement rates were 9.3%, 12.3% and 18.0% on the fitting, evaluation and second-extractor sheets. No ethics review was sought for the annotation, which involved members of the research team labelling text from an already-collected corpus rather than human-subjects research. One annotator is an author, a potential source of bias. Refitting the label-informed correction's parameters θ separately with each annotator's labels gives fitting-disjoint gain sensitivity ranges excluding zero: 0.179 for the author and 0.252 and 0.181 for the other annotators. These analyses assess sensitivity to annotator choice on shared examples, not independent corpus replication.

Data. The private-corpus claims are extracted from public posts collected through X's public API and stored locally. The corpus concerns political discourse and the accounts in it are public and identifiable. Neither the corpus nor the raw annotation sheets are redistributed. The reviewer audit contains only de-identified derived sufficient statistics with pseudonymous rows and no post text or source identifiers. No attempt was made to identify or profile the authors of posts, and no claim in this paper is about any individual or any account. The analysis concerns extraction errors and makes no individual-level claims.

Use. Λ quantifies label information available on a stated benchmark partition. A large value indicates that a score using that partition can perform well. It does not establish negligent benchmark design or invalidate a detector. The diagnostic supports explicit comparison with partition-based priors.

AI USE STATEMENT

Uses requiring disclosure. The research questions, central ideas and all theorems were developed by the authors. We used generative AI tools to implement methods and experiment code, to design and give feedback on experimental protocols and statistical analyses, and to help interpret results. On specific technical problems, these tools suggested alternative approaches and corrections to author-written proofs. They also proposed and assisted in proving supporting mathematical statements, including the comparison $\frac{1}{2}\mathrm{TV} \leq d_{\mathcal{G}}$ after Theorem 2. The authors independently checked all such contributions, with the supplement's computational checks providing additional corroboration. During annotation, annotators used a general-purpose LLM chatbot to translate non-English evidence sentences into English, as described in the Ethics Statement. The private corpus's claims

were extracted by open-weight language models because such extraction is the object of the audit (Appendix S.1). We did not use generative AI tools to generate synthetic datasets, or to produce, revise or adjudicate any label: three human annotators supplied and adjudicated every in-house gold label, with the flagging and translation assistance described in the Ethics Statement.

Other uses. The manuscript was initially drafted by the authors. AI assisted with subsequent revision, principally clarity, language and the organization of existing material, including redrafting passages such as the abstract and adjusting the layout of figures, under the authors' direction. The authors carried out the literature search. AI suggested a few additional related works and checked citation details, and every cited source was checked against its primary publication.

Verification and responsibility. All reported claims were verified against primary sources, behavioral tests, frozen-input replays or independent numerical implementations as appropriate, which is not a substitute for independent human review. Corrected and withdrawn analyses are documented where they affect the reported results (Appendices L and M). Every reported number is emitted from a committed artifact by the verification pipeline, not by a language model. The authors reviewed all AI-assisted work and take full responsibility for the final content, including text, claims and artifacts produced with the aid of generative AI.

REFERENCES

Donald W. K. Andrews and Patrik Guggenberger. Validity of Subsampling and "Plug-in Asymptotic" Inference for Parameters Defined by Moment Inequalities. *Econometric Theory*, 25(3): 669–709, 2009.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal Risk Control. In *International Conference on Learning Representations (ICLR)*, 2024.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset, 2016. arXiv:1611.09268.

Stephen Bates, Anastasios N. Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-Free, Risk-Controlling Prediction Sets. *Journal of the ACM*, 68(6):1–34, 2021. doi: 10.1145/3478535.

Jessa Bekker and Jesse Davis. Learning from Positive and Unlabeled Data: A Survey. *Machine Learning*, 109(4):719–760, 2020. doi: 10.1007/s10994-020-05877-5.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Semi-Supervised Novelty Detection. *Journal of Machine Learning Research*, 11:2973–3009, 2010.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. Translating Embeddings for Modeling Multi-relational Data. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 26, pp. 2787–2795, 2013.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A Large Annotated Corpus for Learning Natural Language Inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642, 2015.

Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. ATOM: Robustifying Out-of-Distribution Detection Using Outlier Mining. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, 2021a.

Lin Chen, Hossein Esfandiari, Gang Fu, Vahab S. Mirrokni, and Qian Yu. Feature Cross Search via Submodular Optimization. In *29th Annual European Symposium on Algorithms (ESA)*, volume 204 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 31:1–31:16, 2021b. doi: 10.4230/LIPIcs.ESA.2021.31.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/Debiased Machine Learning for Treatment and Structural Parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. doi: 10.1111/ectj.12097.

Caroline Choi, Fahim Tajwar, Yoonho Lee, Huaxiu Yao, Ananya Kumar, and Chelsea Finn. Conservative Prediction via Data-Driven Confidence Minimization. *Transactions on Machine Learning Research (TMLR)*, 2024.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased Contrastive Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 8765–8775, 2020.

Stéphan Clémençon and Sylvain Robbiano. Anomaly Ranking as Supervised Bipartite Ranking. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, volume 32 of *PMLR*, pp. 343–351, 2014.

Stéphan Clémençon, Gábor Lugosi, and Nicolas Vayatis. Ranking and Empirical Minimization of U-statistics. *The Annals of Statistics*, 36(2):844–874, 2008.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. EMNIST: Extending MNIST to Handwritten Letters. In *International Joint Conference on Neural Networks (IJCNN)*, pp. 2921–2926, 2017. doi: 10.1109/IJCNN.2017.7966217.

Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The Benchmark Lottery, 2021. arXiv:2107.07002.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Convolutional 2D Knowledge Graph Embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 1811–1818, 2018. doi: 10.1609/aaai.v32i1.11573.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pp. 4171–4186, 2019. doi: 10.18653/v1/N19-1423.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. CLIMATE-FEVER: A Dataset for Verification of Real-World Climate Claims. In *NeurIPS Workshop on Tackling Climate Change with Machine Learning*, 2020.

Zhen Fang, Yixuan Li, Jie Lu, Jiahua Dong, Bo Han, and Feng Liu. Is Out-of-Distribution Detection Learnable? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

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

Daniel Golovin and Andreas Krause. Adaptive Submodularity: Theory and Applications in Active Learning and Stochastic Optimization. *Journal of Artificial Intelligence Research*, 42:427–486, 2011.

Tanya Goyal and Greg Durrett. Annotating and Modeling Fine-grained Factuality in Summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1449–1462, 2021. doi: 10.18653/v1/2021.naacl-main.114.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of NAACL-HLT*, pp. 107–112, 2018.

Michael Gutmann and Aapo Hyvärinen. Noise-Contrastive Estimation: A New Estimation Principle for Unnormalized Statistical Models. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 297–304, 2010.

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 (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep Anomaly Detection with Outlier Exposure. In *International Conference on Learning Representations (ICLR)*, 2019.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching Machines to Read and Comprehend. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, 2015.

Yannis Kalantidis, Mert Bulent Sariyildiz, Noé Pion, Philippe Weinzaepfel, and Diane Larlus. Hard Negative Mixing for Contrastive Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 21798–21809, 2020.

Richard M. Karp. Reducibility among Combinatorial Problems. In *Complexity of Computer Computations*, pp. 85–103. Plenum Press, 1972. doi: 10.1007/978-1-4684-2001-2_9.

Naveen Karunanayake, Suranga Seneviratne, and Sanjay Chawla. CRAFT: Class Ranking Aware Fine-Tuning for Enhanced Out-of-Distribution Detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4119–4128, 2025.

Alexander S. Kechris. *Classical Descriptive Set Theory*, volume 156 of *Graduate Texts in Mathematics*. Springer, 1995.

Tushar Khot, Ashish Sabharwal, and Peter Clark. SciTail: A Textual Entailment Dataset from Science Question Answering. In *AAAI Conference on Artificial Intelligence*, 2018.

Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto, 2009.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the Factual Consistency of Abstractive Text Summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9332–9346, 2020. doi: 10.18653/v1/2020.emnlp-main.750.

Long Lan, Zhaohui Hu, He Li, Tongliang Liu, and Xinwang Liu. C-WOE: Clustering for Out-of-Distribution Detection Learning With Wild Outlier Exposure. *IEEE Transactions on Image Processing*, 35:2066–2079, 2026. doi: 10.1109/TIP.2026.3662593.

Ya Le and Xuan Yang. Tiny ImageNet Visual Recognition Challenge. Technical report, Stanford University, CS 231N, 2015.

Charline Le Lan and Laurent Dinh. Perfect Density Models Cannot Guarantee Anomaly Detection. *Entropy*, 23(12):1690, 2021. doi: 10.3390/e23121690.

Yucen Lily Li, Daohan Lu, Polina Kirichenko, Shikai Qiu, Tim G. J. Rudner, C. Bayan Bruss, and Andrew Gordon Wilson. Position: Supervised Classifiers Answer the Wrong Questions for OOD Detection. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 81689–81713, 2025. URL `https://proceedings.mlr.press/v267/li25ec.html`.

Xiangyu Liu, Yang Liu, and Wei Hu. Knowledge Graph Error Detection with Contrastive Confidence Adaption. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 8824–8831, 2024. doi: 10.1609/aaai.v38i8.28729.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*, 2019.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, 2019.

Aditya Krishna Menon and Robert C. Williamson. Bipartite Ranking: A Risk-Theoretic Perspective. *Journal of Machine Learning Research*, 17(195):1–102, 2016.

George A. Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11): 39–41, 1995. doi: 10.1145/219717.219748.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gülçehre, and Bing Xiang. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 280–290, 2016. doi: 10.18653/v1/K16-1028.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1797–1807, 2018. doi: 10.18653/v1/D18-1206.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading Digits in Natural Images with Unsupervised Feature Learning. In *NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4885–4901, 2020.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10862–10878, 2024. doi: 10.18653/v1/2024.acl-long.585. URL https://aclanthology.org/2024.acl-long.585/.

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.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for Factuality Metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4812–4829, 2021. doi: 10.18653/v1/2021.naacl-main.383.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis Only Baselines in Natural Language Inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics (*SEM)*, pp. 180–191, 2018.

Giovanni Puccetti and Ruodu Wang. Extremal Dependence Concepts. *Statistical Science*, 30(4): 485–517, 2015. doi: 10.1214/15-STS525.

Mark D. Reid and Robert C. Williamson. Information, Divergence and Risk for Binary Experiments. *Journal of Machine Learning Research*, 12:731–817, 2011.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive Learning with Hard Negative Samples. In *International Conference on Learning Representations (ICLR)*, 2021.

Ludger Rüschendorf. On the Distributional Transform, Sklar's Theorem, and the Empirical Copula Process. *Journal of Statistical Planning and Inference*, 139(11):3921–3927, 2009.

Tara Safavi and Danai Koutra. CoDEx: A Comprehensive Knowledge Graph Completion Benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8328–8350, 2020. doi: 10.18653/v1/2020.emnlp-main.669.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An Adversarial Winograd Schema Challenge at Scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 8732–8740, 2020.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. AVeriTeC: A Dataset for Real-world Claim Verification with Evidence from the Web. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. Towards Debiasing Fact Verification Models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pp. 3419–3425, 2019.

Tal Schuster, Adam Fisch, and Regina Barzilay. Get Your Vitamin C! Robust Fact Verification with Contrastive Evidence. In *Proceedings of NAACL-HLT*, pp. 624–643, 2021.

Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct Importance Estimation with Model Selection and its Application to Covariate Shift Adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 20, 2007.

Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. Revisiting DocRED – Addressing the False Negative Problem in Relation Extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8472–8487, 2022. doi: 10.18653/v1/2022.emnlp-main.580.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of NAACL-HLT*, pp. 809–819, 2018.

Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pp. 142–147, 2003. doi: 10.3115/1119176.1119195.

Kristina Toutanova and Danqi Chen. Observed versus Latent Features for Knowledge Base and Text Inference. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, pp. 57–66, 2015. doi: 10.18653/v1/W15-4007.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009. doi: 10.1007/b13794.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *International Conference on Learning Representations (ICLR)*, 2019a.

Qizhou Wang, Bo Han, Yang Liu, Chen Gong, Tongliang Liu, and Jiming Liu. W-DOE: Wasserstein Distribution-Agnostic Outlier Exposure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3530–3545, 2025. doi: 10.1109/TPAMI.2025.3531000.

Zihan Wang, Jingbo Shang, Liyuan Liu, Lihao Lu, Jiacheng Liu, and Jiawei Han. CrossWeigh: Training Named Entity Tagger from Imperfect Annotations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5153–5162, 2019b. doi: 10.18653/v1/D19-1519.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of NAACL-HLT*, pp. 1112–1122, 2018.

Laurin Wischounig, Abdelrahman Abdallah, and Adam Jatowt. Negative Sampling Techniques in Dense Retrieval: A Survey. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 3003–3020, Rabat, Morocco, 2026. doi: 10.18653/v1/2026.findings-eacl.157. URL https://aclanthology.org/2026.findings-eacl.157/.

Jiayang Wu, Wensheng Gan, Jiahao Zhang, and Philip S. Yu. ADKGD: Anomaly Detection in Knowledge Graphs with Dual-Channel Training. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 24(11):1–29, 2025. doi: 10.1145/3748321. arXiv:2501.07078.

An Yang et al. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhen Yang, Ming Ding, Chang Zhou, Hongxia Yang, Jingren Zhou, and Jie Tang. Understanding Negative Sampling in Graph Representation Learning. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, pp. 1666–1676, 2020. doi: 10.1145/3394486.3403218.

Haiyun Yao, Zongbo Han, Huazhu Fu, Xi Peng, Qinghua Hu, and Changqing Zhang. Out-Of-Distribution Detection with Diversification (Provably). In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. DocRED: A Large-Scale Document-Level Relation Extraction Dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 764–777, 2019. doi: 10.18653/v1/P19-1074.

Sergey Zagoruyko and Nikos Komodakis. Wide Residual Networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2016. doi: 10.5244/C.30.87.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 93–104, 2018.

Qinggang Zhang, Junnan Dong, Keyu Duan, Xiao Huang, Yezi Liu, and Linchuan Xu. Contrastive Knowledge Graph Error Detection. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 2590–2599, 2022. doi: 10.1145/3511808.3557264.

A NOTATION AND READING GUIDE

Guide to the appendices.

- **Reading aids.** This appendix fixes the notation, §A.1 interprets every displayed equation, Appendix C.1 separates each claim from its classical ingredient and states its boundary, and Appendix C places the work against the literature in full.
- **Reproducibility scope.** Appendix B states what can be checked from the paper, what the released packages recompute, and what is not redistributable.
- **Theory.** Appendix E states and proves every result and opens with a table mapping each one to the checks that exercise it. Appendix F gives the estimators, intervals and reporting conventions for Λ.
- **Controlled detector interventions.** Appendix Q holds the complete crossed comparisons for both detectors and both graphs, Appendix U the schema-prior dose-response, and Appendix T the language-model and NLI baselines.
- **Human-identified error processes.** Appendix X is the prospectively specified Re-DocRED design, Appendix J the two FRANK error processes, Appendix Y the CoNLL++ null, and Appendix S the per-score result on our own annotated sheet. Appendix R compares each correction with its prior control, and Appendix V gives the transfer analyses across graphs, languages and extractors.
- **Cross-domain studies.** Appendices G and G.1 cover summary grounding errors, Appendices H and H.1 the prospective question-answering study and its post-hoc refit, Appendix I the failed frozen-head transfer, and Appendix Z the image intervention with its protocol amendment.
- **The diagnostic.** Appendix K reports the full survey, Appendix D the VitaminC case, and Appendix O the datasheet protocol and released tool output.
- **Audit trail.** Appendix L records the model-selection chronology, including where selection was not blind, Appendix M the withdrawn held-out-account probe, Appendix N the shifted-split analysis, Appendices P and W the annotation-budget studies, and Appendix AA the simulation detail.

Higher scores always mean more anomalous. Errors are the positive class for AUC. The subscript "pos" in p_{pos} instead retains the conventional meaning of a *correct* claim in contrastive training. We use e and c for cell marginals to avoid confusing these conventions.

Density and score conventions. The training label $L = 1$ denotes a pool draw and $L = 0$ an auxiliary draw. Its prior $\alpha = \Pr(L = 1)$ is separate from error prevalence π. The canonical anomaly score is $d_q = \log(q/p)$ after removing the training-prior constant. An observed score $\widetilde{d}_q = d_q + \zeta$ has the same ordering, but recovering the normalized density requires the integral in Result 1a. Ratios at positive-error/zero-correct density have value $+\infty$, while jointly zero-density sets are ignored. For $A(s, U, V)$, the first law U supplies the positive draw and V the negative draw. The oracle $A^*(U, V)$ is the supremum over measurable scores, not the best observed fitted head.

Local notation in the extensions. Result 1e uses capital $P, E, C, Q_{\pm}$ for probability laws and lower-case $r_{\pm} = dQ_{\pm}/dP$ for their density ratios. Its $z = \eta - \pi$ is a centered posterior and Z, Z' are independent scalar copies under P. Neither is the tilt normalizer $Z(\theta)$ or the sensitivity multiplier z. Result 4b uses β for failure probability and $r(k, n_1, n_0, \beta)$ for a confidence radius. Result 5c uses Δ for the difference of two log-score perturbations and $\tau > 0$ for a margin threshold. The map-family protocol uses m maps and B_{perm} permutations. These quantities are local to the named results. Result 1f uses u for an inspection probability, b for its expected pool fraction, h for its cutoff and γ for acceptance at a cutoff tie. Result 5d uses $A(s, U, V)$ for AUC with positive law U and negative law V, d_{TV} for error-to-auxiliary distance, $d_{\mathcal{G}} \le d_{\mathrm{TV}}$ for its ranking-class refinement (the form stated as Theorem 2) and ϵ_{ctr} for unrestricted population contrastive AUC shortfall. $\delta(x)$ remains the log-density error throughout the regret analysis.

In Appendix G, $\widehat{P}, \widehat{Q}_m, \widehat{Q}_g$ are empirical fitting laws, $d(i)$ is a source donor index, $z(x)$ is a standardized feature vector, and λ is the auxiliary dilution fraction, not the ridge penalty. The ridge coefficient is written explicitly as 0.01 there.

Symbol	Meaning
$X, x,\ Y$	Random item, its realization, error indicator (1 error, 0 correct).
P, E, C, Q, μ	Pool, error, correct-item and auxiliary probability laws, with common reference measure μ.
$p, p_{\mathrm{err}}, p_{\mathrm{pos}}, q$	Corresponding densities relative to μ, at the stated resolution.
$\pi,\ \eta(x)$	Error prevalence $\Pr(Y = 1)$, posterior error probability $\Pr(Y = 1 \mid X = x)$.
$\rho(x)$	Likelihood ratio $p_{\mathrm{err}}(x)/p_{\mathrm{pos}}(x)$, posterior odds are $\pi\rho/(1-\pi)$.
q_0, q_θ	Original corruption density and its fitted exponential tilt.
ψ	Information retained by an analysis or score. Distributional statements can be applied to $\psi(X)$.
$\ell_q, d_q,\ d^*, d_\theta$	Contrastive logit and its anomaly orientation, oracle and modelled anomaly scores.
$T, k,\ a, b$	Cell map and its number of cells, cell indices.
e_a, c_a	Cell probabilities conditional on error and correctness, respectively. Both vectors sum to one.
s_a, h_{ab}	Score in cell a, pair comparison with one-half credit for a tie.
$D_{ab},\ w_{ab}$	Signed pair asymmetry $e_a c_b - e_b c_a$, ordered pair mass $e_a c_b$.
S_{ab}	Sign of D_{ab}, with zero diagonal, used in the sensitivity gradients.
$\Lambda(T)$	Maximum cell-only AUC minus $1/2$, for the specified distributions and map.
$t_a^{(j)}$	Cell ratio $\widehat{e}_a^{(j)}/(\widehat{e}_a^{(j)} + \widehat{c}_a^{(j)})$ fitted on cross-fitting half j, the score whose held-out AUC the cross-fitted value measures (Appendix F).
$\Gamma_{j\to k}$	One fitting-to-evaluation direction of that estimate, before averaging over directions and repetitions.
φ, θ, d	Fixed feature vector, fitted parameter vector and their dimension.
$Z(\theta), \widehat{Z}(\theta), \lambda$	Population normalizer, finite-sample normalizer, nonnegative ridge coefficient.
$\overline{\varphi}_{\mathrm{err}}$	Weighted average of features on annotated errors used for fitting.
$\delta(x), \varepsilon$	Log-density error $\log(q_\theta/p_{\mathrm{err}})$ and its supremum absolute value.
B, ε_B	Retained subset of the analysis space and local supremum error on it.
$n_y, W_y, n_{\mathrm{eff},y}$	Number, total weight and Kish effective count of observations with label y.
R, z	Number of repeated cross-fitting splits, multiplier for an analytic sensitivity interval (Appendix F).
ϕ, χ	A randomized test, taking values in $[0, 1]$, and its conditional expectation given the posterior quantile U. Both are test functions, never divergences.
$\chi^2(Q\|P)$	Chi-squared divergence $\int (dQ/dP - 1)^2 dP$, unrelated to the test function χ above.
ϑ	A parameter value inside the natural domain $\{\vartheta : Z(\vartheta) < \infty\}$, used where θ is held fixed.
I_d	Identity matrix of dimension d.
M, Δ, τ	Oracle margin $d^*(X^+) - d^*(X^-)$, difference of two log-score perturbations, and a margin threshold.
u_e, u_c	Error mass in cells with no correct mass, and correct mass in cells with no error mass.
$E_{<j}, C_{<j}$	Prefix sums of e and c over cells preceding j in ratio order.
$O(k \log k)$	Time complexity in the number of cells k, not an asymptotic statistical statement.

Table 2: Core notation. Hats denote sample estimates. Local quantities used only in a derivation are defined where they enter. All logarithms are natural.

The finite-pool remark after Result 1f uses N for the number of items in the pool and K for an integer inspection count, neither of which is the information constant K of Result 1g below. Result 1g uses D_σ for the two joint item/label laws, a for their posterior separation, κ for the oracle AUC advantage, K for one-label KL divergence and n for the number of i.i.d. annotated pairs. Its adaptive form uses H_n^σ for the law of the complete query history and K_{ad} for the largest single-observation divergence. The atomless form of Result 1d uses F_η for the distribution function of η, L_k for its level sets of positive mass, W for the within-level tie breaker, U for the resulting uniform variable and g for the quantile function of a likelihood-ratio law. Results 5e–5f use G, g, p_g, π_g for a stratum, its value, mass and error prevalence, ω_g for conditional comparison weights, and P_0 for a training pool that can differ from target pool P. These quantities are local to those results.

The finite form of Result 1d uses m_i for item masses, L_ℓ for equal-posterior item sets, M_ℓ for their masses and S_ℓ for cumulative masses. $\mathcal{B}$ contains those cumulative boundaries and their reflections, t_j are its sorted cut points, ω_j the intervening masses, and $\mathcal{J}_\ell$ the interval indices inside level ℓ. $L_{\ell j}$ are the item subsets required by the mass-partition criterion and v_j the constructed ratio values. These symbols are local to that criterion.

The corollary to Result 1d uses $\mathcal{T}(P, Q)$ for the binary testing region and F for a criterion determined by that region. In Appendix H.1, G_f, G_t count fitting and evaluation source components, b indexes bootstrap replicates, and M_{bg}^f, M_{bg}^t are their sampled component multiplicities. The refitted score is s_b, $g(i)$ maps a row to its source component, h_i is its generator/source stratum, and w_{bi} is its evaluation multiplicity. These bootstrap quantities are separate from the population inspection budget b and the theoretical stratum G used elsewhere.

Operators, functions and abbreviations. EMMA (Error-Model Matched Auditing) is the name we give the label-informed correction of Appendix S.1, and "EMMA-corrected" marks a score to which it has been added. PMI is pointwise mutual information, used as a partial-input feature in the benchmark survey. $\mathrm{TV}(U, V)$ is total variation distance and d_{TV} its value for a stated pair. KL is Kullback–Leibler divergence, $\mathrm{kl}(u, v)$ its Bernoulli special case $u \log(u/v) + (1-u)\log[(1-u)/(1-v)]$, and χ^2 the chi-squared divergence. $A(s, U, V)$ is AUC, the area under the receiver-operating-characteristic (ROC) curve, with positive law U and negative law V. We write $\mathrm{AUC}(s)$ for the same quantity whenever the pair of laws is the one the surrounding text fixes, which for an audit is the error and correct-claim pair. A^* is its supremum over measurable scores, and $\mathcal{R}$ the resulting regret. $d_{\mathcal{G}}(E, Q) = \sup_v |A(v, E, P) - A(v, Q, P)|$ is the ranking discrepancy of Theorem 2, with $\frac{1}{2}\mathrm{TV}(E, Q) \le d_{\mathcal{G}}(E, Q) \le \mathrm{TV}(E, Q)$. $\mathrm{Prec}_b(u)$ is the precision of an inspection rule u spending pool-mass budget b. Var is variance and sign the sign function. $C_{\mathrm{FP}}, C_{\mathrm{FN}}$ are false-positive and false-negative costs in a testing risk. SE is a standard error, MDE_{80} the effect a design can detect with 80% power, and z_u the standard-normal u quantile. Subscripts name the object rather than an index in $\widehat{\Lambda}_{\mathrm{cf}}$ (cross-fitted), $\widehat{Q}_{\mathrm{bal,perm}}$ (balanced permutation) and ϵ_{ctr} (contrastive). Hats denote sample estimates, $\mathbb{E}$ expectation with its law in the subscript, and Pr probability. All logarithms are natural.

Resolution. A cell is a group of items sharing a selected feature value. A relation map is coarser than a head-type–relation–tail-type signature. Lexical maps may use selected content. "Cell-only" means that the score uses only the retained feature, not that the feature is content-free. A distribution of $\psi(X)$, sometimes called a push-forward, assigns probability to the feature values induced by drawing X. A theorem about that distribution applies to scores measurable on those feature values. For pool recovery, support coverage means $P \ll Q$, or $q > 0$ P-almost surely. In contrast, $Q \ll P$ excludes auxiliary mass outside the pool and appears in the positive regret theorem. These are different directions of absolute continuity. The ratio q/p is defined on the pool even if it can be zero. Finite log-fidelity bounds additionally require compatible support of q_θ and p_{err} and finite log ratios on the evaluated support. These are population assumptions. Coverage alone is not a complete statistical consistency theorem for a fitted neural model.

The four Λ quantities and their interpretation. The paper uses one symbol for a population functional and two estimators of it, and only some of them bound anything. The convention is rigid throughout.

Term	Meaning	May be called a ceiling?
Population $\Lambda(e, c)$	the identity of Result 4, on the stated map	yes, this is the theoretical ceiling
Plug-in $\widehat{\Lambda}_{\mathrm{plug}}$	$\Lambda(\widehat{e}, \widehat{c})$, exact for the supplied rows	yes, but only for those rows
Cross-fitted $\widehat{\Lambda}_{\mathrm{cf}}$	a pairwise sensitivity statistic: out-of-fold, attenuated by uncertain pair signs, and a held-out AUC only under the fitting-coverage condition of Appendix F	**no**
Result 4b interval	population confidence statement under its i.i.d. assumptions	yes, when those assumptions hold

A cross-fitted value is a sensitivity estimate, and adding $\frac{1}{2}$ to it gives neither an empirical nor a population maximum. Where the splits cover every evaluated cell, it gives the mean held-out AUC of a fitted cell prior. A reported row carries that interpretation only when fitting coverage is established, and it lists the covered and total direction counts and a flag for any display truncation. Wherever this paper writes "ceiling" the quantity beside it is the population identity or the plug-in, never the cross-fitted estimate.

Metrics and uncertainty. AUC includes half credit for ties. P@5% is precision among the highest-scored items at the stated budget. CAGED/ADKGD and their priors use five percent of the pre-injection graph size, while CCA uses five percent of its full ranked pool (Appendix Q). Other studies state their own budget and tie convention. AUC regret is the oracle's AUC minus the evaluated score's AUC. Our human-labelled sheets use nominal stratum calibration weights because their realized sampling probabilities are unavailable. Their bootstrap ranges and alternative-weight analyses are sensitivity summaries, not confidence intervals with established design coverage. Kish's effective count summarizes concentration of weights. Equal precision to an unweighted sample would require additional assumptions about outcomes and sampling.

Experimental language. "Published" denotes the upstream corruption configuration. "Within-type" denotes replacements observed in the same relation slot. These are synthetic, slot-attested errors, rather than human-observed extraction errors. **The two names denote constructions, not properties shared by every row.** Our re-implementation of the within-type construction skips an anchor with no acceptable replacement, so 100% of its injected FB15k-237 rows are slot-attested. The detectors' patched generator instead falls back to the published generator after 30 rejected draws, at a rate estimated only by the diagnostic replay of Appendix Q. The published configuration replaces a uniformly drawn slot with a uniformly drawn filler, and half its anomalies are wholly random triples. It therefore inserts a slot-attested argument in only 2.1% of rows on FB15k-237 and 15.0% on WN18RR. Read as a *schema* property rather than an attestation property, the distinction is weaker still. The published arm's heads carry a type the relation is never observed with in 68.7% of rows on FB15k-237, and WN18RR's relation names carry no schema for the notion to refer to. A crossed 2×2 separates the corruption used to train a detector from the error process used to evaluate it. Compare columns within the same evaluation row. A partial-input baseline is a model using a restricted input, whereas a Λ ceiling refers to all scores on one specified finite map, which may retain less information.

A.1 Interpretation of the displayed equations

The numbered equations are explained below in the order of the main argument, followed by the theoretical extensions, correction and estimators.

Main text.

- (1) contrasts the two rankings the paper is about. A contrastive model trained to separate the pool from a corruption ranks items by q/p, the corruption density over the pool density. An audit wants items ranked by $p_{\mathrm{err}}/p_{\mathrm{pos}}$, the error density over the correct-claim density.

The corruption q appears in the first and not in the second, and this asymmetry is the central difficulty.

- (2) fixes the generative picture. The pool p is the mixture of correct claims and errors in proportion $1-\pi$ to π, the ratio ρ compares the two class-conditional densities at an item, and the last equality says the posterior odds of error are ρ rescaled by the prior odds. Ranking by ρ and ranking by the posterior are therefore the same ranking.
- (3) is the central theorem. The first equality says the two corruptions pose the same binary testing problem against the pool, in the strongest sense that every achievable pair of error probabilities is shared. The next two say their induced audit rankings achieve the best and the worst possible AUC against the same fixed errors, A^* and $1-A^*$.
- (4) is the positive counterpart. Audit regret $\mathcal{R}(s)$, the AUC an oracle attains minus the AUC of the evaluated score, is bounded by two terms divided by the correct-claim proportion $1-\pi$. The first, $2d_{\mathcal{G}}(E,Q)$, at most $2\mathrm{TV}(E,Q)$, is infidelity: how far the corruption is from the real error law, as seen by rankings against the pool. The second, $\epsilon_{\mathrm{ctr}}(s)$, is contrastive shortfall: how far the score is from solving its own training problem. Small values of both, relative to $1-\pi$, therefore *suffice* for small regret. Neither is necessary: by Result 1 any $q \propto p\,h(\rho)$ with h positive and strictly increasing induces the oracle ordering, however far it lies from E.
- (5) defines the diagnostic. D_{ab} is the signed asymmetry of a pair of cells, positive when errors favour cell a relative to correct items. Λ is half the total absolute asymmetry, and it equals the largest advantage over chance that any score using only the cell can achieve.

Theory appendix.

- (6) is the finite feasibility criterion. Reading it as an instruction: the cut points split the unit interval into pieces of mass ω_j, and each posterior level must be able to hand out its items into groups whose masses match exactly the pieces lying inside it. Items cannot be divided, which is what makes the problem combinatorial and, already with two posterior levels, NP-complete in the weak sense.
- (7) builds the explicit pair on an arbitrary pool: both corruptions have density ratio one plus or minus a small multiple t of the centered posterior $z=\eta-\pi$, so they are mirror images around the pool itself.
- (8) and (9) verify that the pair is indeed matched on the summaries most often used to describe difficulty. Both corruptions give the same optimal contrastive AUC, the same total variation distance to the pool and the same chi-squared divergence, with Z and Z' independent copies of z. These are specified summaries only, which is why the atomless construction is needed for the complete testing region.
- (10) converts a ranking into an inspection policy. For a rule u that inspects each item with some probability and spends a budget b of pool mass, the precision of the inspected set is the posterior-weighted mass it captures divided by the budget. Maximizing precision at a fixed budget therefore means concentrating u where η is largest.
- (11) and (12) are the two lines of the finite-annotation lower bound. The first states it in total-variation form and the second converts that to the information K by Pinsker's inequality. It is a lower bound, not a guarantee, and its quantifiers matter. Two worlds indexed by σ differ only in the direction of the posterior separation a. Whatever score $\widehat{s}$ an auditor produces from n annotated pairs, *at least one* of the two worlds leaves its *expected* audit regret at or above $\kappa \max\{0, 1-\sqrt{nK/2}\}$. Here κ is the oracle advantage over chance, and K is the information one annotated pair carries about which world is in force. The expectation is over the annotations and any randomness in $\widehat{s}$. The bound is vacuous once n exceeds about $2/K$, which marks where this argument stops forcing regret rather than a sample size at which any particular method succeeds. The first line shows the source: the auditor must effectively distinguish the two worlds, so its regret is controlled by the total variation between the two n-sample laws.
- (13) writes the AUC of a cell-only score as a double sum over ordered cell pairs, weighted by $e_a c_b$, with h_{ab} giving full credit for a correctly ordered pair and half for a tie. The identity $h_{ab}+h_{ba}=1$ is what makes reversing a score send AUC A to $1-A$.

- (14) is the computational form of Λ. After sorting cells by their ratio, prefix sums $E_{<j}$ and $C_{<j}$ turn the double sum into one pass, which is why the cost is $O(k \log k)$ and dominated by the sort.
- (15) says Λ moves by at most the total variation distance by which its two input distributions move, which is what makes a finite-sample bound possible at all.
- (16), (17) and (18) give that bound. The first defines the per-class radius $b(k, n, \beta)$, which comes from a uniform deviation inequality for a k-cell distribution estimated from n draws with failure probability at most $\beta/2$ per class. The second adds the two class radii. The third is the coverage statement the sum yields for the plug-in estimate.
- (19) is the support floor. If a fraction u_e of the error mass lies in cells that contain no correct items, then the indicator of those cells alone already achieves $u_e/2$ above chance, so Λ is at least that. The same holds with the classes exchanged, hence the maximum.
- (20) is the margin bound. M is the oracle's margin between a random error and a random correct item, and ε bounds the log-fidelity error of the model, so pairs can only be lost when the oracle margin is smaller than twice that error.
- (21) restricts the same argument to a retained region B, paying the mass of each class outside it, which is what allows a bound when the error is large only on a rare set.
- (22) tightens the bound by removing pairs that fall in the same cell, which cannot be lost because the surrogate score is constant there.
- (23) and (24) replace the supremum error by an average one. Pairs are split at a margin threshold τ, and Markov's inequality converts the remaining probability into mean absolute log error divided by τ.
- (25) is the appendix statement of the bound displayed in the main text.
- (26) is the identity on which the bound rests. Scoring errors against the entire pool, rather than against correct items only, shrinks the advantage over chance by the factor $1 - \pi$. The reason is that an error drawn against another independent error receives mean comparison credit $1/2$ by exchangeability, whether or not individual scores tie.
- (27) applies the bound inside each stratum g and averages with the comparison weights ω_g, so fidelity is only required within a group rather than across groups.
- (28) adds the price of a shifted pool. When the score was fitted against a training pool P_0 and is applied to a target pool P, the distance between the two pools enters the bound exactly like infidelity does.
- (29) characterizes optimal cell scores: a score is AUC-optimal exactly when it orders each pair the way the pair masses $w_{ab} = e_a c_b$ do, and pairs with equal masses may be ordered freely. This freedom is what Result 7a exploits.

Derived correction.

- (33) says what fitting a tilt changes about the ranking. The corrected anomaly score is the original log density ratio d_0 plus a linear term in the features, $\theta^\top \varphi$, minus a constant $\log Z(\theta)$. Because the constant does not depend on the item, ranking with the correction only requires adding $\theta^\top \varphi(x)$.

Estimation appendix.

- (30) is the cross-fitted estimator. Each pair's sign is taken from one half of each class and its magnitude from the other half, the two roles are exchanged, and the result is averaged over R repeated splits. This avoids using the same noise to choose and evaluate a pair's sign. Incorrect signs can attenuate the estimate, and the procedure is not generally unbiased.
- (31) is the condition under which that bridge holds: every cell carrying evaluation mass must have some fitting mass, so the ratio that scores it is defined. Only the total matters, not the split between classes.
- (32) is the delta-method variance used for the diagnostic interval. Each class contributes the variance of its own influence terms $g_{e,a}$ and $g_{c,a}$ divided by its sample size, which is why a class with few labelled items dominates the width.

B REPRODUCIBILITY SCOPE

Verification from the paper. Appendix E gives complete proofs and maps every proved result to the executable checks that exercise it on exact or exhaustive instances (Table 5). Appendix F specifies the estimators and their intervals, and Appendix L documents the model-selection chronology. Prospectively specified designs, the protocol amendments, the failed criteria and the withdrawn analyses are identified where they occur rather than collected in a separate section. Every reported empirical result, table entry, figure value and in-text result is generated from committed artifacts and checked against the manuscript by the verification pipeline, which fails if the text and the artifact disagree. Seeds, splits and settings are recorded in each artifact and in the appendix that reports it, and a registered reproduction gate re-derives every artifact on its registered submission list and compares it byte for byte. The nine experiments the gate declares as development are outside that list and are not reproduced by it.

Independent recomputation. The standalone Λ tool and its documentation are available at `https://anonymous.4open.science/r/lambda-index-212A`, and Appendix O reproduces its datasheet output. The supplement holds eight self-contained packages, each with a verifier that exits non-zero on any disagreement.

- *Proof checks* runs every check in Table 5 and reports them per result.
- *Detectors* recomputes every CAGED, ADKGD and CCA cell from per-seed P@5%, the CCA interaction and its exact sign test, the CAGED–ADKGD reversal, the schema-type prior's paired differences from both detectors on their own FB15k-237 pools (with a canonical checksum of each seed's pool, so a pool rebuilt with the upstream loader can be compared), and the fitting-disjoint correction result from pseudonymous frozen scores. It also includes executable CAGED/ADKGD and CCA intervention patches, the paired CCA benchmark builder, its pinned anomaly draw, public-source revisions and setup instructions. These expose the training interventions without distributing upstream models or the CCA pretraining checkpoint.
- *Provenance* holds the CCA protocol and benchmark checksums and both frozen error models, and recomputes the models' content hashes.
- The *Summary*, *QA* and *Data2txt* packages recompute frozen-score statistics from source-pair matrices, including all controls, primary intervals and development planning gates, without checkpoints or raw text.
- A seventh package replays every post-hoc *QA refit*, with independent pair calculations and selected checks under a second optimizer.
- An eighth package, *FRANK*, replays all bootstrap statistics from numeric features and saved heads, checks AUC by independent pair enumeration, and refits the original heads and a subset of bootstrap heads by default, with an option to refit them all. It has a new anonymous release manifest, distinct from the unchanged original chronology records. This verifies feature-to-result computation, not raw-text preparation or feature extraction.

Material that is not redistributed. The private corpus and the raw annotation sheets are not released, for the reasons given in the Ethics Statement, and the packages carry pseudonymous rows with no post text and no source identifiers. No package retrains the upstream detectors without their own data, so those comparisons are auditable as calculations rather than re-executable end to end. We state this limit rather than describe the packages as a full reproduction of every experiment.

C RELATED WORK, EXTENDED

Recent objective-misalignment and auxiliary-shift work. Li et al. (2025) argue that supervised uncertainty and features can answer questions different from OOD detection, and discuss limitations of Outlier Exposure among possible interventions. Our formal question fixes the representation and asks which latent error rankings are compatible with a contaminated pool even at an optimal contrastive score. Wang et al. (2025, Theorem 1) already bound OOD risk using auxiliary-to-target Wasserstein discrepancy and statistical complexity terms, and enlarge auxiliary coverage through implicit synthesis. Result 5d therefore does not establish the first connection between auxiliary fidelity and risk. Its specific object is a population AUC-regret transfer inequality for tie-aware audits

inside a fixed contaminated mixture, with a ranking-discrepancy term bounded by total variation and unrestricted contrastive AUC shortfall. It does not replace their finite-sample guarantee. C-WOE reweights mixed ID/OOD auxiliary data (Lan et al., 2026). That addresses contamination of the auxiliary, while our identification question concerns the latent decomposition of the observed pool. These distinctions delimit our statements rather than assert that adjacent methods never handle contaminated data.

Clean references, identifiability and learnability. When clean correct items are available, contrasting the pool against them solves the audit problem. Blanchard et al. (2010) reduce novelty detection with a contaminated unlabelled sample to binary classification against nominal data, show that the resulting detectors are optimal whatever the novelty distribution, and prove the decomposition identifiable under a minimal-proportion condition on the novelties (their Proposition 5). In our notation, flagging items that are unlike the clean reference C ranks by the pool-to-reference ratio $p/p_{\mathrm{pos}} = (1-\pi)+\pi\rho$, which is increasing in ρ and so recovers the oracle ordering. Contrastive auditors replace C with a synthetic corruption, and Theorem 1 shows that the difficulty of that contrast then cannot certify alignment with the error ranking. Fang et al. (2022) prove that out-of-distribution detection is not learnable in some settings, a statement about learning a detector from finite samples. Our limit concerns population identification with the target held fixed, and holds with unlimited data.

Limits of accurate density models. Two lines reach conclusions adjacent to ours. Le Lan & Dinh (2021) show that a perfect density model does not guarantee anomaly detection, because likelihood is not invariant under reparametrization, so likelihood-based detection rests on implicit assumptions that ought to be stated. On compact support, Clémençon & Robbiano (2014) show that anomaly ranking against a uniform reference distribution *is* bipartite ranking, which is the sense in which a chosen reference fixes the ranking target. These results concern the interpretation of density-based anomaly rankings. They do not state the fixed-target, equal-testing-region reversal studied here. Our obstruction survives with the representation held fixed. On a fixed cell map, with the observed pool recovered exactly and the error prevalence known, the decomposition of that pool into errors and correct claims is still not determined, so the audit ranking is not either (Results 1a and 1b).

Choice of auxiliary negatives. That the choice of auxiliary negatives matters is established in out-of-distribution detection, and we do not claim it. Outlier Exposure studies the auxiliary set's characteristics (Hendrycks et al., 2019). ATOM mines informative outliers after arguing that many are useless or harmful (Chen et al., 2021a). Data-Driven Confidence Minimization proves guarantees when auxiliary unknowns resemble those met at test time (Choi et al., 2024). Diversification work frames performance through auxiliary-to-test shift (Yao et al., 2024). CRAFT reports exposure bias toward test distributions resembling the exposed outliers, which motivates its outlier-free fine-tuning method (Karunanayake et al., 2025). Those findings already establish that auxiliary composition and similarity matter. Hard-negative sampling and mixing (Robinson et al., 2021; Kalantidis et al., 2020) raise difficulty to improve learned representations. Result 1d concerns a different object: difficulty, even complete testing difficulty, does not certify alignment with a latent error ranking. Our question is whether observations (p, q, π) identify a latent error ranking at all, even when the contrastive optimum is known. Positive-unlabelled (PU) non-identifiability motivates the unknown-target witness (Results 1b–1c), in which the decomposition of the pool varies. It does not state the fixed-target experiment of Theorem 1 (Result 1d), which keeps the latent target fixed while varying q at identical binary-testing difficulty. Result 1d makes that separation explicit using classical equivalences, and the crossed interventions examine its practical consequence. We claim neither the discovery of exposure bias nor a new binary-classification identity. The image study matches a restricted learner's measured AUC and does not reproduce the theorem's full testing-region equivalence.

Two further lines provide context: Confident Learning (Northcutt et al., 2021) identifies likely label errors, and adaptive submodularity gives acquisition guarantees when its structural assumptions hold (Golovin & Krause, 2011). We do not establish those assumptions for our annotation allocation. Distribution-free risk control (Bates et al., 2021; Angelopoulos et al., 2024) provides calibrated prediction guarantees under its own sampling and loss assumptions, and Appendix W records why we do not claim one.

Density-ratio estimation and the estimand distinction. NCE asks whether the data distribution p can be recovered by discriminating it from a chosen noise distribution q. With q known and covering the support, the population density-ratio score identifies p, so changing q need not cost consistency for density estimation (Gutmann & Hyvärinen, 2010). An audit makes a different use of the same object: it deploys that score directly as a ranking of which observed items are errors. The target ranking is determined by $p_{\mathrm{err}}/p_{\mathrm{pos}}$, while the contrastive ranking is determined by q/p. Knowing q makes reconstruction of p possible, but ranking by the learned ratio does not perform that reconstruction. Consistency for the first task therefore does not identify the second.

This is also the boundary with the negative-sampling literature. That q changes a contrastive objective is known, and that literature studies the choice through optimization, through how negatives are mined, and through false-negative mitigation (Yang et al., 2020; Wischounig et al., 2026). Our claim is not that q enters contrastive learning. It is that once the resulting score is interpreted as evidence of *real* error, q becomes part of the population functional being ranked. Result 1a isolates the distinction: in the controlled construction the pool is recovered exactly for every q, while the audit ranking induced by that same optimal score moves with q. A contrastive estimator can be consistent for the distribution it was asked to estimate and inconsistent for the auditing interpretation placed on its score.

Role of each experiment. The controlled density-ratio construction holds the pool and real error process fixed. Changing q changes the optimal score's audit AUC in all 300 instances, illustrating the difference between density recovery and audit ranking. Where the error process is known, the matched corruption recovers oracle ordering (Appendix AA). These numerical checks illustrate the proved statements rather than replace their assumptions. The benchmark pairs test whether selected maps retain label information after artifact removal. HANS (McCoy et al., 2019) gives a constant-feature zero, FEVER-Symmetric (Schuster et al., 2019) gives a measured collapse, and ANLI (Nie et al., 2020) fails the prespecified monotonicity prediction (Appendix F.1). The detector interventions test sensitivity to synthetic error construction. The human-label experiments instead test fitted score corrections: gains over the base persist on fitting-disjoint rows, but prior controls match them, and a smoothed prior has higher mean AUC on Re-DocRED. The fitting-disjoint split was not blind in model-selection chronology (Appendix L). These experiments support different claims.

Dependence on the private corpus. The cross-benchmark diagnostic and survey use published labels. The crossed detector comparisons, ranking inversion and prespecified CCA experiment use public graphs and injected errors, without the human sheets' nominal weights. Re-DocRED, CoNLL++ and FRANK also use external labels. Claims about the human-annotated correction and frozen cross-extractor transfer rely on our private corpus. Independence from that corpus does not remove each experiment's own support, construction or generalization limits.

Per-seed detail of the ranking reversal. The shared lineage that makes CAGED and ADKGD weak as *independent* evidence about corruption design is what makes this comparison well controlled. Because ADKGD forks CAGED's corruption code, both detectors receive the same injected anomalies at a given seed. The pairing is therefore exact, and no matching assumption is needed. Beyond the unanimity reported in §4.1, every seed's paired margin on within-type errors is positive, ranging from 0.0027 to 0.0190, around a mean of 0.008. This is consistency across the four observed runs, with limited evidence about stability beyond them. On WN18RR the order does not reverse and ADKGD leads on both error processes.

C.1 Contribution boundaries and evidence requirements

The central result concerns what contrastive observations identify about a latent error ranking. It is not a claim that auxiliary choice, objective misalignment, likelihood-ratio ranking or distribution-shift bounds are new. The following table distinguishes the mathematical witnesses from their classical ingredients and from the experimental claims.

Scope of each line of evidence. The theory and the experiments answer different questions, and no experiment in this paper instantiates Result 1d: equalizing a measured discrimination proxy is not equalizing the population testing region, and we do not claim it is. Table 4 states, for each line

Claim	Contribution in this paper	Boundary
Identification	Fixed pool and prevalence admit every measurable score ordering (1b–1c), with a finite-label bound for separated worlds under i.i.d. or adaptive annotation (1g).	Mixture non-identifiability and two-point testing are inherited. Ordering is modulo null sets and represented by a measurable real score.
Equal hardness	A fixed latent target permits opposite audit orders at identical complete testing difficulty on any atomless pool, at any continuous difficulty level (1d), or three specified summaries on arbitrary pools (1e).	Reindexing and the distributional transform are classical. Finite feasibility has an exact mass-partition criterion, reducing to symmetric profiles for distinct posteriors. With ties, rational-input feasibility is NP-complete. Result 1e equates summaries only.
Inspection budget	The same witnesses maximize and minimize precision at every fixed pool-mass budget (1f).	Posterior threshold optimality is classical. Atomic ties require randomized inspection.
Positive guarantee	Error-to-auxiliary ranking discrepancy $d_{\mathcal{G}}$ (at most TV) and unrestricted contrastive AUC shortfall control audit regret, with conditional and pool-shift extensions (5d–5f).	Bounded-loss transfer is inherited. Unknown population terms do not certify fitted detectors.
Diagnostic	A stated-map population ceiling, support witness and reporting protocol expose construction dependence.	The categorical maximum AUC is inherited from Chen et al. (2021b).
Generalization	Crossed detector interventions and separate human-error and image evaluations test consequences. FRANK's two human error processes give a positive interaction with refitted heads. QA passes its fixed-head strong criterion, with the generic contrast unresolved after refitting. Unchanged-head Data2txt transfer fails.	CoNLL++ is unresolved. FRANK's full crossover fails, one row carrying the effect. Image and Summary miss their strong criteria. QA does not outperform supervision or test new generator families.

Table 3: Claim boundaries. Mathematical existence results do not supply a label-free auxiliary-selection algorithm or a universal empirical advantage.

of evidence, what is held fixed, what varies, where the uncertainty comes from, and the claim it supports.

Evidence	Target	What varies	Uncertainty	Supports
Result 1d	fixed latent target	the auxiliary Q	theorem	hardness does not identify audit fidelity
CAGED (§4.1)	synthetic errors	corruption	four seeds	corruption-dependent fitted conclusions
CCA (§4.1)	synthetic errors	joint matched intervention	five seeds	independent architectural replication
Re-DocRED (App. X)	human removals	corruption reweighting	document bootstrap	matching can help on a real error process
FRANK (App. J)	two human error types	label-free auxiliary	refit bootstrap	asymmetric corruption–error interaction
QA and Data2txt (App. H, I)	human-annotated hallucinations	matched against generic	source bootstrap	limited positive generalization, failed transfer

Table 4: Evidence scope. Each row names the claim its line of evidence supports. The theorem is not instantiated by any experiment, and the experimental rows are about fitted detectors under a stated intervention, not about population testing regions.

Estimands of the interventions. The controlled detector comparisons estimate the effect of changing the specified training-corruption procedure while keeping an evaluation error process fixed. They do not show that a fitted neural score equals the population density ratio, or separate every change in feature distribution, discrimination difficulty and semantic content caused by that intervention. The QA comparison additionally holds empirical generator/source marginals fixed and uses within-stratum pairs. It still does not establish equal hardness or isolate a single semantic mechanism. Its original fixed-head generic contrast passes, while the post-hoc refit interval includes zero. The failed no-refitting transfer is a separate scope limitation.

Annotation and comparison budgets. The matched auxiliaries use fitting error labels. Supervised controls use labels from the same fitting rows. The generic head does not use those labels to select its auxiliary, although the common development design uses annotations. Consequently, a matched-versus-generic gain measures the value of this annotation-informed construction relative to that comparator. It establishes neither label-free auditing nor lower annotation cost than supervised learning. The stronger supervised and prior controls limit the interpretation of positive matched-versus-generic contrasts.

Practical implications of the guarantees. The population discrepancy and unrestricted contrastive shortfall in Results 5d–5f are not estimated by fitting loss, a sample AUC or the distance between two raw empirical supports. Thus those bounds organize the sufficient conditions rather than certify a deployed detector. The operational consequence is to state the target error process, compare corruption choices on the same held-out target errors, and include controls given the same fitting annotations. The recorded calculation audits check implementation and arithmetic at the declared scope. They do not recreate annotation or prove independence of sources, generators or pretraining data.

D VitaminC: label balance and the measured contrast

VitaminC (Schuster et al., 2021) pairs claims with near-identical supporting and refuting evidence. In the analyzed data, 69.1% of claims have both *supports* and *refutes* rows, and 68.9% have equal counts of those labels. Our reported $\widehat{\Lambda}_{\mathrm{cf}} = 0.0389$, however, uses *not enough information* (NEI) versus the remaining labels. Balancing *supports* against *refutes* does not balance their union against

NEI. The paired benchmark comparison fixes the feature scheme but uses each benchmark's maximizing label contrast, so it does not hold the binary task fixed.

For claim-only cells, exact per-claim balance in the measured binary contrast would imply equal class marginals and zero Λ. Balance in a different contrast does not imply this. Claims with only one of the support/refute labels illustrate another source of unequal cell marginals, but we do not quantify their contribution to the reported NEI-versus-rest value. The residual diagnostic therefore does not by itself establish failure of VitaminC's paired-evidence design.

E Complete statements and proofs

This appendix states the population identities, finite-cell diagnostic and regret bounds. In the main text, Result 1 is Proposition 1, Result 1d is Theorem 1, Result 5d is Theorem 2, and Results 3 and 4 are Propositions 2 and 3. Every other result is stated only here and keeps its identifier. Numerical evaluations are separate from the proofs. Densities are taken relative to a common reference measure μ at the chosen resolution, with the support conditions in Appendix A. Zero-mass cells may be omitted. Log-ratio bounds are used only where their values are finite.

Table 5 lists every result in this appendix with its proof idea, the classical ingredient it uses and the machine checks that exercise it. Each check is an exact rational, symbolic or exhaustive computation written independently of the proof steps, and all run in every reproduction gate and in the supplement's proof-check package. They test the stated identities, inequalities and constructions on concrete instances, including ties, zero class mass, boundary cases and every counterexample quoted in the text. A universal statement still rests on its written proof, since no finite check establishes it. A generator verifies that every result heading here has an entry and that every listed check exists.

E.1 Contrastive estimation and audit identification

Result 1: population score. Let $L = 1$ indicate a sample from p, $L = 0$ a sample from q, and $\alpha = \Pr(L = 1) \in (0, 1)$ be the training class proportion, distinct from error prevalence π. Bayes' rule gives

$$\log \frac{\Pr(L = 1 \mid x)}{\Pr(L = 0 \mid x)} = \log \frac{p(x)}{q(x)} + \log \frac{\alpha}{1 - \alpha}.$$

The second term is constant and has no ranking effect. Reversing the logit gives $d_q = \log(q/p)$ up to a constant. Also

$$\eta(x) = \frac{\pi p_{\mathrm{err}}(x)}{p(x)} = \frac{\pi \rho(x)}{(1 - \pi) + \pi \rho(x)}.$$

For $q = p_{\mathrm{err}}$, $q/p = \eta/\pi$ has the target order. More generally, $q = p\, h(\rho)/C_h$, with $C_h = \int p(x) h(\rho(x))\, d\mu(x) \in (0, \infty)$ and positive strictly increasing h defined on the range of ρ, yields $q/p = h(\rho)/C_h$ and the same conclusion. If $\rho = +\infty$ occurs on positive pool mass, h must be defined there and still satisfy $C_h < \infty$.

Scope. This is the population density-ratio identity underlying NCE (Gutmann & Hyvärinen, 2010).

Result 1a: recovering the pool does not recover the audit target. Assume $q > 0$ P-almost surely, so the corruption covers the pool support. For a known corruption and its exact optimal anomaly score,

$$p(x) = q(x) e^{-d_q(x)}.$$

For an observed score $\widetilde{d}_q = d_q + \zeta$ with unknown finite constant ζ, the scale-free recovery formula is

$$p(x) = \frac{q(x) e^{-\widetilde{d}_q(x)}}{\int q(u) e^{-\widetilde{d}_q(u)}\, d\mu(u)}.$$

Here μ is the common reference measure and the denominator is $e^{-\zeta}$. Thus an additive constant is irrelevant to ordering but must be normalized away to recover density values. Without coverage, an infinite logit on $q = 0$ does not identify the pool density there, so this recovery statement does not

Result	Proof idea	Classical ingredient	Machine checks
Result 1	Bayes rule for the population classifier	NCE optimum	random pools, four increasing transforms (2)
Result 1a	normalizing qe^{-d_q}	none	random pools, derivative sign of the $\sqrt{\rho}$ example (2)
Result 1b	explicit perturbation of the pool	mixture non-identifiability	random constructions, every order with ties (3)
Result 1c	bounded monotone transform of the score	none	extreme finite scores, continuous pool with unbounded score (2)
Result 1d (finite, uniform)	bijection of tests under permutation	equivalence of binary experiments	all permutations to five cells, all priors and costs (4)
Corollary to 1d	equal regions, opposite AUCs	none	via the region and AUC checks of 1d (2)
Result 1d (atomless)	distributional transform with a measurable tie breaker	distributional transform, measure isomorphism	symbolic pool with tied and continuous posteriors, grid integration, budgets (5)
Result 1d (finite, ties)	region determines the ratio law, cut-point partition	Neyman–Pearson	all weak orders on 774 small pools with ties, symmetric and asymmetric profiles (6)
Complexity corollary	reduction from Partition	NP-completeness of Partition	subset enumeration on every weight vector to length six (1)
Result 1e	likelihood-ratio AUC identity with ties	likelihood-ratio optimality	exact pools with ties and zero class mass, accuracy identity (3)
Result 1f	threshold exchange argument	Neyman–Pearson	exhaustive feasible vertices, symbolic pool (2)
Minimax consequence	constant score and degenerate decomposition	none	every score on random pools (1)
Result 1g	two-point testing, KL chain rule, Pinsker	Le Cam two-point method	exact risks to six labels, every adaptive policy to six queries (2)
Results 2 and 3	conditioning on both cells	none	random marginals and scores (1)
Result 4	pairwise maximization, prefix sums	categorical maximum AUC	every weak order of the cells, weighted, quadratic definition, exhaustive held-out ratio order (3)
Result 4a	nested score classes	none	random refinements (1)
Result 4b	Lipschitz bound, Hoeffding, union bound	Hoeffding inequality	simplex boundary, exact binomial coverage (2)
Result 3a	indicator of one-class cells	none	random counts, weighted floor below Λ (2)
Result 5	pairs lose credit only inside the margin band	none	random pools with exact regret (1)
Result 5a	charge pairs outside B in full	none	trimmed bound on the same instances (1)
Result 5b	guaranteed diagonal ties	none	wide attainment search, stated counterexample (3)
Result 5c	threshold split and Markov inequality	Markov inequality	ties and score shifts (1)
Result 5d	bounded kernel transfer, mixture AUC identity	none	all three-cell orderings, exact $d_{\mathcal{G}}$, matching law (3)
Result 5e	per-stratum Result 5d	none	imperfect shared scores, balanced oracle (2)
Balanced-permutation control	exact expectation over permutations	none	every within-stratum permutation (1)
Result 5f	kernel triangle inequality	none	independent pair sums, stated counterexample (2)
Result 6	convex exponential family, moment equation	exponential families	finite differences, convexity, moment matching, known tilts (5)
Results 7 and 7a	pairwise optimality of the likelihood-ratio order	bipartite ranking optimality	random pools, the equal-ratio example (2)

Table 5: Proof map. The count in parentheses is the number of independent checks registered for the result. Each runs in the reproduction gate, and the supplement's proof-check package runs all of them per result.

apply. Thus the pool is recoverable although ranking by d_q still depends on q. *The pool denominator matters.* If $q \propto p_{\mathrm{pos}}\sqrt{\rho}$, then

$$\frac{q}{p} \propto \frac{\sqrt{\rho}}{(1-\pi)+\pi\rho}.$$

The derivative has the sign of $(1-\pi)-\pi\rho$, so the expression increases and then decreases, turning at $\rho = (1-\pi)/\pi$. Its decontaminated counterpart $q/p_{\mathrm{pos}} \propto \sqrt{\rho}$ is increasing throughout. Normalizing constants are suppressed only because they do not affect either ordering.

Result 1b: every strict finite-cell ordering is compatible with one pool. Fix $p_i > 0$ with $\sum_i p_i = 1$ and a known $\pi \in (0,1)$. Choose distinct real values z_i in any desired order and center them so $\sum_i p_i z_i = 0$. For sufficiently small $t > 0$, let $a_i = 1 + tz_i$ satisfy $0 < a_i < 1/\pi$, and define

$$p_{\mathrm{err}}(i) = p_i a_i, \qquad p_{\mathrm{pos}}(i) = \frac{p_i(1-\pi a_i)}{1-\pi}.$$

Both are positive probability distributions because $\sum_i p_i a_i = 1$. Their mixture is p, and

$$\rho_i = \frac{(1-\pi)a_i}{1-\pi a_i}$$

is strictly increasing in a_i. The observable (p, q, π) therefore admits every strict cell ordering. For example, at $\pi = 1/2$, the pool $(0.65, 0.35)$ arises from correct and error laws $(0.5, 0.5)$ and $(0.8, 0.2)$, or from these components exchanged. The audit orders are opposite in the two decompositions.

Checks. Numerical checks cover 2000 constructions with 0 failures. Random alternative decompositions reverse an order in 76% of draws. These checks corroborate the construction, rather than establish it. Once p_{err} is known, $a_i = p_{\mathrm{err}}(i)/p_i$ is fixed and the displayed formula preserves its order for every compatible prevalence. Finite annotations estimate this additional distribution. They do not supply population knowledge exactly.

Result 1c: every measurable population ordering is compatible with one pool. Let P be a probability measure, $\pi \in (0,1)$, and r a measurable real-valued score, finite P-almost surely. Put $h(u) = u/(1+|u|)$, bounded and strictly increasing, and $z = h(r) - \mathbb{E}_P[h(r)]$. This expectation exists even when r has no finite mean. If $B = \|z\|_\infty > 0$, choose $0 < t < \min(1, \pi^{-1} - 1)/B$ and $a = 1 + tz$, so that $\mathbb{E}_P[a] = 1$ and $0 < a < 1/\pi$ almost surely. Define $dE/dP = a$ and $dC/dP = (1-\pi a)/(1-\pi)$. Both are positive normalized densities with $P = \pi E + (1-\pi)C$, and

$$\frac{dE}{dC} = \frac{(1-\pi)a}{1-\pi a},$$

whose derivative in a is $(1-\pi)/(1-\pi a)^2 > 0$. Since $t > 0$ and h is strictly increasing, this ratio has exactly r's strict comparisons and ties. If $B = 0$ then r is almost surely constant and $E = C = P$ preserves its ties.

Scope. All statements are modulo P-null sets. This holds for discrete, continuous and mixed P, with no Lebesgue density or nonatomic assumption, and realizes orderings representable by measurable scores rather than arbitrary order relations. Relative to Result 1b it replaces finite sums by expectations and bounds a monotone transform. This extends the domain of Result 1b without a separate proof technique.

Result 1d: identical training difficulty, opposite audit orderings. Fix $n \geq 2$, $p_i = 1/n$, $\pi \in (0,1)$, and a strictly positive decomposition e, c of p whose ratios e_i/c_i are all distinct. Let v be any strictly positive probability vector with distinct coordinates, and for a permutation σ set $q_\sigma(i) = v_{\sigma(i)}$ with score $s_\sigma = q_\sigma/p$. Then the binary testing region for (p, q_σ) does not depend on σ, the scores s_σ realize all $n!$ strict orderings, and their audit AUCs include both $\mathrm{AUC}^* = \mathrm{AUC}(e/c, e, c) > 1/2$ and $1 - \mathrm{AUC}^*$. The decomposition is held fixed while q varies. This is a different operation from Result 1b, which varies the decomposition.

Training equivalence. For a randomized test $\phi_i \in [0,1]$ accepting the auxiliary class with that probability in cell i, the two class-conditional acceptance probabilities are $(\sum_i p_i\phi_i,\ \sum_i q_\sigma(i)\phi_i)$.

Setting $\psi_j = \phi_{\sigma^{-1}(j)}$ and using uniformity of p makes this $(\sum_j p_j\psi_j,\ \sum_j v_j\psi_j)$, and the map is a bijection between tests. The entire achievable region is identical, not merely its upper envelope, so any optimal decision risk depending only on training class and action agrees at every prior and cost. Explicitly, for auxiliary-class prior $\nu \in (0,1)$ and nonnegative false-positive and false-negative costs $C_{\mathrm{FP}}, C_{\mathrm{FN}}$, $\sum_i \min\{(1-\nu)C_{\mathrm{FP}}p_i,\ \nu C_{\mathrm{FN}}q_\sigma(i)\}$ is permutation invariant. Likewise $D_f(q_\sigma\|p) = \sum_i p_i f(q_\sigma(i)/p_i)$ is unchanged for every f-divergence for which it is defined.

Audit claim. Distinct coordinates of v realize every strict ordering. Ordering by e_i/c_i maximizes pairwise AUC, since each unordered pair contributes $\max(e_ic_j, e_jc_i)$, and reversing it minimizes every pair. As the two tie-aware ranking kernels sum to one, $\mathrm{AUC}(s) + \mathrm{AUC}(-s) = 1$. Distinct target ratios and positive masses give $\mathrm{AUC}^* > 1/2$. A suitable decomposition exists for any π by the finite case of Result 1c.

Scope. Reindexing equivalent binary experiments is classical, and the correspondence among ROC curves, Bayes risks and divergences is standard (Reid & Williamson, 2011). The derivation above is an explicit bijection and asserts none of those equivalences as new. The consequence here is their insufficiency for ordering a fixed latent error component inside the observed pool. The q_σ are different observable distributions: the claim is not that training samples are identically distributed in fixed coordinates, but that training-discrimination summaries invariant under the reindexing carry no information about alignment with the audit target. The construction does not preserve feature-dependent losses, geometric hardness definitions, optimization trajectories, or risk within a hypothesis class not closed under the permutation, and it requires p uniform. For non-uniform p, arbitrary permutations need not preserve the training problem, which a test verifies. The atomless form below removes uniformity. Exact-arithmetic checks cover all permutations through five cells for the testing-region identity, with independent pairwise AUC recomputation.

Corollary to Result 1d: testing-region criteria cannot certify alignment. Let $\mathcal{T}(P,Q)$ denote the achievable binary testing region defined above, and let F be any function of that region. In either fixed-target construction of Result 1d, finite uniform or atomless, choose Q_+ and Q_- whose density-ratio scores respectively attain the oracle and reversed audit orderings. Then

$$F(\mathcal{T}(P,Q_+)) = F(\mathcal{T}(P,Q_-)),$$
$$A(dQ_+/dP, E, C) = A^* > \tfrac{1}{2}, \qquad A(dQ_-/dP, E, C) = 1 - A^* < \tfrac{1}{2}.$$

Thus F cannot determine even the sign of advantage over chance on this pair. In particular, a certificate determined only by the testing region cannot validly certify positive alignment for the first member without making the same, false certification for the second. It may always abstain, but it cannot resolve this pair. The proof is immediate from equality of the regions and the two audit AUCs in Result 1d. Optimal ROC curves, binary Bayes discrimination risks and the specified divergence summaries all agree on the pair.

Scope. This is a consequence of the existing witness, not an additional impossibility assumption or a claim that Q_+ and Q_- are indistinguishable as distributions in fixed coordinates. Criteria using semantic structure or additional target information need not be functions of $\mathcal{T}$ and are outside the corollary.

Result 1d on atomless pools: every continuous difficulty level. Let the item space be standard Borel and let $P = \pi E + (1-\pi)C$ be atomless, with $0 < \pi < 1$, posterior $\eta = \pi\, dE/dP$ not P-almost surely constant, and $A^* = \sup_s A(s, E, C)$. Let $g : (0,1) \to (0,\infty)$ be strictly increasing with $\int_0^1 g = 1$. There is a measurable U, uniform on $(0,1)$ under P and nondecreasing in η, such that the auxiliaries

$$\frac{dQ_+}{dP} = g(U), \qquad \frac{dQ_-}{dP} = g(1-U)$$

satisfy (i) $\mathcal{T}(P,Q_+) = \mathcal{T}(P,Q_-)$, both equal to the testing region of Lebesgue measure against $g\,du$ on $(0,1)$, (ii) $A(dQ_+/dP, E, C) = A^* > \frac{1}{2}$ and $A(dQ_-/dP, E, C) = 1 - A^*$, and (iii) at every inspection budget $b \in (0,1)$, inspecting the top b of the pool under dQ_+/dP maximizes the precision of Result 1f and under dQ_-/dP minimizes it. Every binary experiment (P', Q') with $Q' \ll P'$ whose likelihood ratio has an atomless law under P' has the same region as (i) with g its quantile function, because a dominated region is determined by the law of its likelihood ratio (see the finite condition below). Equal complete difficulty at any such level therefore leaves the audit sign unidentified.

Construction of U. Let F_η be the distribution function of η under P and $F_\eta(t-)$ its left limit. The level sets $L_k = \{\eta = c_k\}$ of positive mass are countable. Each normalized restriction of P to L_k is an atomless Borel probability on a standard Borel space, so by the measure isomorphism theorem (Kechris, 1995, Theorem 17.41) there is a Borel $W_k : L_k \to [0,1]$ pushing it to Lebesgue measure. Put $W = W_k$ on L_k, $W = 0$ elsewhere, and

$$U = F_\eta(\eta-) + W\,\{F_\eta(\eta) - F_\eta(\eta-)\}.$$

Given η, W is uniform on each L_k and irrelevant elsewhere, so U has the law of the distributional transform and is uniform (Rüschendorf, 2009). If $\eta(x) < \eta(x')$ then $U(x) \le F_\eta(\eta(x)) \le F_\eta(\eta(x')-) \le U(x')$.

Proof. (i) For a randomized test ϕ let $\chi = \mathbb{E}_P[\phi \mid U]$, a measurable function of U with values in $[0,1]$. Its class probabilities under (P, Q_+) are $(\int_0^1 \chi, \int_0^1 \chi g)$, and every measurable $\chi : (0,1) \to [0,1]$ arises from the test $\chi(U)$. Under (P, Q_-) the pair is $(\int_0^1 \chi, \int_0^1 \chi(u) g(1-u)\,du)$, and $v = 1-u$ maps one set of test functions bijectively onto the other. (ii) U is atomless under P, hence under $E, C \ll P$, so independent $X^+ \sim E$ and $X^- \sim C$ tie in U with probability zero. When $\eta(X^+) \neq \eta(X^-)$, U orders the pair as η does, and η is an optimal score. When both lie in one L_k, dE/dP and dC/dP are constant there, so the pair is conditionally i.i.d. from P restricted to L_k and $U(X^+) > U(X^-)$ with probability $\frac{1}{2}$, matching the tie credit. Common values of η carrying no mass have probability zero. Hence $A(g(U), E, C) = A(\eta, E, C) = A^*$. As g is strictly increasing, $g(1-U)$ is a strictly decreasing function of $g(U)$, and tie-aware AUCs of reversed scores sum to one. $A^* > \frac{1}{2}$ because η is not constant. (iii) $\{U > 1-b\}$ has pool mass b, contains $\{\eta > h\}$ and lies in $\{\eta \ge h\}$ for the cutoff level h. The optimality argument of Result 1f needs only this, not how the tied level is split, because η is constant there. $\{U < b\}$ is the corresponding minimizer.

Result 1d on finite pools: an exact criterion including posterior ties. Let the pool put mass $m_i > 0$ on item $i \in \{1, \dots, k\}$, with $\sum_i m_i = 1$, posterior error probabilities $\eta_i \in [0,1]$, and $\pi = \sum_i m_i \eta_i \in (0,1)$. The error and correct laws have masses $e_i = m_i \eta_i / \pi$ and $c_i = m_i(1-\eta_i)/(1-\pi)$. Let $L_1, \dots, L_r$ be the sets of items at the distinct posterior levels, listed in increasing order. Put

$$M_\ell = \sum_{i \in L_\ell} m_i, \qquad S_0 = 0, \quad S_\ell = \sum_{h=1}^{\ell} M_h, \qquad \mathcal{B} = \{S_\ell, 1 - S_\ell : 0 \le \ell \le r\}.$$

Write the distinct elements of $\mathcal{B}$ in increasing order as $t_0, \dots, t_J$, where $t_0 = 0$ and $t_J = 1$, and define interval masses $\omega_j = t_j - t_{j-1} > 0$. These cut points refine both the increasing and decreasing posterior-level mass intervals. Let $\mathcal{J}_\ell = \{j : S_{\ell-1} \le t_{j-1} < t_j \le S_\ell\}$.

There exist auxiliaries $Q_+, Q_- \ll P$ with identical complete testing regions whose density-ratio scores attain A^* and $1 - A^*$ if and only if, for every ℓ, the items in L_ℓ admit a partition into sets $(L_{\ell j})_{j \in \mathcal{J}_\ell}$ such that

$$\sum_{i \in L_{\ell j}} m_i = \omega_j \quad \text{for every } j \in \mathcal{J}_\ell. \tag{6}$$

Items are indivisible: this condition permits grouping items with the same posterior, but not splitting their pool masses. For $r = 1$ the target is constant, the condition holds trivially and both audit AUCs equal $1/2$.

Necessity. A dominated binary testing region determines, and is determined by, the pool-law of its likelihood ratio (Reid & Williamson, 2011). On a finite pool its upper boundary is the Neyman–Pearson curve, with slopes equal to ratio values and horizontal extents equal to their pool masses. Let the two ratio scores have the same such law. By (29), an optimal score strictly orders items with unequal posteriors and may order a posterior tie arbitrarily. Reversing a score maps its AUC a to $1-a$, so a score attaining $1 - A^*$ is exactly the reversal of such a score and strictly reverses unequal posteriors. Thus every S_ℓ is a cumulative mass boundary of the increasing score's law. Every $1 - S_\ell$ is a boundary of the decreasing score's law, hence also of their common law. Grouping consecutive ratio levels between the cut points in $\mathcal{B}$ produces sets of mass ω_j. No item is split and no set crosses a posterior-level boundary, yielding (6).

Sufficiency and construction. Reflection preserves $\mathcal{B}$, so $\omega_j = \omega_{J+1-j}$. Choose positive, strictly increasing values $v_1, \dots, v_J$ satisfying $\sum_j \omega_j v_j = 1$, for example $v_j = j / \sum_{h=1}^{J} h\omega_h$. For $i \in L_{\ell j}$

set

$$\frac{dQ_+}{dP}(i) = v_j, \qquad \frac{dQ_-}{dP}(i) = v_{J+1-j}.$$

Both ratios have the same pool-law, assigning mass ω_j to v_j, and both integrate to one. Across unequal posterior levels they strictly increase and strictly decrease, respectively. They therefore attain the oracle and reversed audit AUCs. Their score ties stay inside posterior levels, so randomized inspection at a cutoff also attains the maximum and minimum precision of Result 1f at every pool-mass budget.

Distinct-posterior corollary. If every L_ℓ is a singleton, no additional cut can split its mass. The criterion reduces to $\{S_\ell\}_\ell = \{1 - S_\ell\}_\ell$, equivalently $m_i = m_{k+1-i}$ for all i when items are listed by increasing posterior. This recovers the symmetric-profile criterion and its uniform-pool special case. With ties the criterion is strictly more general. For masses $(1/4, 1/4, 1/2)$ and posteriors $(1/4, 1/4, 3/4)$, the ratios $(2/3, 2/3, 4/3)$ and $(4/3, 4/3, 2/3)$ share a testing region and have audit AUCs $3/4$ and $1/4$. The individual mass profile is not symmetric in any posterior-respecting order. Conversely, masses $(2/5, 3/10, 3/10)$ with a lower posterior on the first item require splitting the two higher-posterior items into masses $1/5$ and $2/5$, which is impossible. Posterior ties alone do not ensure feasibility.

Computational corollary: exact finite feasibility is weakly NP-complete. Given positive rational masses and rational posteriors in binary encoding, deciding whether the pair in Result 1d exists is NP-complete, even with two posterior levels. Membership in NP follows from (6): a certificate assigns each item to one of at most $2r - 1$ intervals, and its level membership and rational mass sums can be checked in polynomial time. For hardness, take positive integers $a_1, \ldots, a_n$ from PARTITION, with $W = \sum_i a_i$. Create one item with mass $1/3$ and posterior $1/4$, and n items with masses $2a_i/(3W)$ and posterior $3/4$. The prevalence is $7/12$, so this is an admissible fixed target. The cut set is $\{0, 1/3, 2/3, 1\}$. The lower-posterior item fills one interval, while the higher-posterior items must split into two sets of mass $1/3$. This is equivalent to a subset of the integers summing to $W/2$. The construction has polynomial encoding length, completing the reduction from the classical NP-complete problem (Karp, 1972). This concerns exact auxiliary-pair existence with the target supplied. It is not a hardness result for neural training, approximation, or label acquisition. The reduction establishes NP-hardness in the ordinary sense only. With two posterior levels, after scaling to integer masses, feasibility is a subset-sum instance, so it admits a pseudo-polynomial dynamic program and is not strongly NP-hard unless $\mathrm{P} = \mathrm{NP}$. Whether feasibility is strongly NP-hard when the number of posterior levels is unrestricted is left open.

Scope. The distributional transform and measure isomorphism theorem are cited above. Increasing and decreasing rearrangements are also classical (Puccetti & Wang, 2015). The contribution is the consequence for latent audits: complete binary testing information does not identify alignment with the error ranking on any atomless pool, at any continuous difficulty level. The construction uses η, so it is an existence result and not a label-free selection rule. It does not preserve feature geometry, optimization or a restricted model class. Exact symbolic checks use a non-uniform pool whose posterior has two tied levels and a continuous region and is not monotone in the item. They verify uniformity of U, equal Neyman–Pearson curves at two difficulty levels, audit AUCs A^* and $1 - A^*$ and the budget extremes, with an independent grid integration. Finite checks compare the mass-partition criterion with every score weak order on small rational pools, including ties. They independently verify the convex hulls of deterministic tests, audit AUCs and budget extrema. Integer subset enumeration checks the computational reduction. These checks exercise the proofs and do not establish their universal claims. The earlier special case $g(u) = 1 + \delta(2u - 1)$ with $p(x) = 2x$, $dE/dP = 1 + \eta(2x^2 - 1)$ and $\pi = \frac{1}{2}$, where $U = x^2$, is checked separately by independent integration.

Result 1e: equal AUC, total variation and Pearson divergence for any pool. Fix a probability law $P = \pi E + (1 - \pi)C$, with $0 < \pi < 1$. E and C are respectively the error and correct-item laws and are held fixed. No uniformity, continuity or finite-support assumption is needed. Write $a = dE/dP$, which exists because $E \ll P$, and choose the version satisfying $0 \le a \le 1/\pi$. The posterior error probability is $\eta = \pi a$. Define

$$z(x) = \eta(x) - \pi, \qquad Z = z(X), \quad Z' = z(X'), \qquad X, X' \overset{\text{iid}}{\sim} P.$$

Thus Z, Z' are scalar transformed scores, not items sampled from the pool. They satisfy $\mathbb{E}Z = 0$ and $|Z| \le \max(\pi, 1-\pi)$. For $0 < t < 1/\max(\pi, 1-\pi)$ define probability laws Q_+, Q_- by

$$r_\pm(x) := \frac{dQ_\pm}{dP}(x) = 1 \pm tz(x). \tag{7}$$

Both derivatives are strictly positive and integrate to one. On a common density space this says $q_\pm = p(1 \pm tz)$. Since η is strictly increasing in the extended likelihood ratio $a/(dC/dP)$, with ratio $+\infty$ and posterior $\eta = 1$ where only E has mass, r_+ has the oracle comparisons and ties, and r_- reverses every strict comparison. Consequently their audit AUCs are A^* and $1 - A^*$, where $A^* = \sup_s \mathrm{AUC}(s, E, C)$. If $E = C$, both scores are constant and $A^* = 1/2$.

Define $\mathrm{AUC}^*_{\mathrm{ctr}}(P, Q) = \sup_s \mathrm{AUC}(s, Q, P)$: auxiliary draws are the positive class in this training comparison, whereas errors are the positive class in the audit comparison. Then

$$\mathrm{AUC}^*_{\mathrm{ctr}}(P, Q_\pm) = \tfrac{1}{2} + \tfrac{t}{4}\mathbb{E}|Z - Z'|, \tag{8}$$

$$\mathrm{TV}(Q_\pm, P) = \tfrac{t}{2}\mathbb{E}|Z|, \qquad \chi^2(Q_\pm \| P) = t^2\mathbb{E}Z^2. \tag{9}$$

Here $\mathrm{TV}(Q, P) = \frac{1}{2}\int |dQ/dP - 1|\, dP$ and $\chi^2(Q\|P) = \int (dQ/dP - 1)^2\, dP$ for $Q \ll P$. Equal-prior optimal classification accuracy is $[1 + \mathrm{TV}(Q, P)]/2$, so it also agrees for the two auxiliaries.

Proof of the training AUC identity, including ties. For any density ratio $r = dQ/dP$, put $H(u, v) = \mathbf{1}[u > v] + \frac{1}{2}\mathbf{1}[u = v]$. Likelihood-ratio optimality gives $A = \mathbb{E}[r(X)H(r(X), r(X'))]$, with expectation under $P \otimes P$. Exchange X, X' and use $H(u, v) + H(v, u) = 1$ to obtain

$$2A = \mathbb{E}\max\{r(X), r(X')\} = 1 + \tfrac{1}{2}\mathbb{E}|r(X) - r(X')|.$$

Substitute $r = r_\pm$ to prove (8). The two distance identities follow directly by substitution, and the audit reversal follows because the two comparison kernels sum to one even at ties.

Scope. This is an existence result using the latent posterior, not an implementable way to select auxiliaries without error information. It covers pools with atoms, where the complete-region form of Result 1d can fail. It equates the three specified population summaries, not the complete testing region, other f-divergences, finite-model performance or optimization. For example, unequal third moments of Z can separate other divergence summaries. Result 1d retains the stronger testing-region equivalence under its own assumptions.

Result 1f: equal hardness and opposite inspection-budget optima. Retain $P = \pi E + (1-\pi)C$ and the auxiliaries of Result 1e. An inspection rule $u : \mathcal{X} \to [0, 1]$ gives the probability of inspecting each item. For an expected pool fraction $b \in (0, 1)$, require $\mathbb{E}_P u = b$ and define

$$\mathrm{Prec}_b(u) = \frac{\pi\mathbb{E}_E u}{b} = \frac{\mathbb{E}_P[\eta u]}{b}, \qquad \eta = \pi\frac{dE}{dP}. \tag{10}$$

This is the error probability conditional on inspection, not an assertion about the expected ratio from a finite random number of inspections.

Proof. Choose a threshold h satisfying $P(\eta > h) \le b \le P(\eta \ge h)$ and put

$$u^*(x) = \mathbf{1}[\eta(x) > h] + \gamma\mathbf{1}[\eta(x) = h], \qquad \gamma = \frac{b - P(\eta > h)}{P(\eta = h)}$$

when $P(\eta = h) > 0$. When that mass is zero, set $\gamma = 0$. Such a threshold exists because $\eta \in [0, 1]$. For any feasible u,

$$b\{\mathrm{Prec}_b(u^*) - \mathrm{Prec}_b(u)\} = \mathbb{E}_P[(\eta - h)(u^* - u)] \ge 0.$$

The constant term cancels because both rules have budget b. The integrand is nonnegative both above and below the threshold and zero on its boundary. Thresholding the smallest posterior values instead gives the minimum by the same argument with the sign reversed. Thus thresholding $r_+ = dQ_+/dP$ from the top maximizes precision and thresholding $r_- = dQ_-/dP$ from the top minimizes it, for every budget b, while their training-hardness summaries agree. Ties need not produce a strict performance gap. The maximum and minimum precision do not in general sum to one, unlike reversed-score AUCs.

Finite pools. For the uniform law on a fixed pool of N items and an integer inspection count K, take $b = K/N$ and select exactly K items, randomizing uniformly within a cutoff tie if necessary. Formula (10) then equals expected precision at that fixed count. A reported P@5% additionally uses its declared rounding and tie convention. This connects the population ordering to budgeted auditing without identifying a fitted neural score with the population ratio.

Minimax consequence of non-identification. For fixed P, Q and known $0 < \pi < 1$, let $\mathcal{D}(P,\pi) = \{(E,C) : P = \pi E + (1-\pi)C\}$. An auditor $\mathcal{A}$ using only these observables and independent randomness outputs a measurable score. Then

$$\sup_{\mathcal{A}} \inf_{(E,C)\in\mathcal{D}(P,\pi)} \mathbb{E}_{\mathcal{A}} A(\mathcal{A}(P,Q,\pi),E,C) = \tfrac{1}{2}.$$

A constant score guarantees one half. Conversely, the admissible decomposition $E = C = P$ gives one half to every score, including randomized scores, by exchanging the independent pair and crediting ties by one half. This unrestricted minimax statement alone is elementary and permits an uninformative target. The following separated two-world result removes that degeneracy and quantifies finite-label regret. It is the more informative decision-theoretic statement.

Result 1g: a separated ambiguity and a finite-annotation lower bound. The identification obstruction need not rely on error and correct laws becoming identical. Fix $0 < \pi < 1$ and $0 < a < \min(\pi, 1-\pi)$. Let the pool P be uniform on $\{-1,+1\}$, and define two joint laws D_σ for an item X and its error indicator Y, indexed by $\sigma \in \{-1,+1\}$, through

$$\Pr_{D_\sigma}(Y = 1 \mid X = x) = \eta_\sigma(x) = \pi + \sigma a x.$$

Both laws have pool P and prevalence π. Their class-conditional masses are

$$E_\sigma(x) = \frac{\pi + \sigma a x}{2\pi}, \qquad C_\sigma(x) = \frac{1-\pi-\sigma a x}{2(1-\pi)}.$$

These are strictly positive probability laws. Write $\kappa = a/[4\pi(1-\pi)] > 0$. The oracle AUC is $1/2 + \kappa$ in both worlds, but its strict ordering reverses between them. Suppose an auditor receives $n \geq 0$ i.i.d. annotated pairs from D_σ and outputs an arbitrary, possibly randomized score $\widehat{s}$. It may additionally know P, π, a fixed auxiliary Q, and unlimited independent samples from P and Q, whose laws are the same in both worlds. Define expected audit regret using fresh independent class draws after fitting. Then

$$\max_{\sigma\in\{-1,+1\}} \mathbb{E}_{D_\sigma^n} \mathcal{R}_\sigma(\widehat{s}) \geq \kappa\{1 - \mathrm{TV}(D_+^n, D_-^n)\} \tag{11}$$

$$\geq \kappa \max\{0, 1 - \sqrt{nK/2}\}, \qquad K = a \log \frac{(\pi+a)(1-\pi+a)}{(\pi-a)(1-\pi-a)}. \tag{12}$$

For $n = 0$, the minimax regret over these two separated worlds is exactly κ, attained by a constant score. Thus unlimited contrastive observations cannot eliminate this positive ambiguity without target-label information.

Proof. Let $v = \mathrm{sign}(s(+1) - s(-1)) \in \{-1, 0, 1\}$. Direct half-tie AUC evaluation gives $A(s, E_\sigma, C_\sigma) = 1/2 + \sigma\kappa v$, so $\mathcal{R}_\sigma(s) = \kappa(1 - \sigma v)$. Map the fitted score to a randomized test choosing world $+$ with probability $(1+v)/2$, conditional also on any fitting randomness. Its two error probabilities sum to at least $1 - \mathrm{TV}(D_+^n, D_-^n)$: integrating the smaller of the two sample likelihoods gives that minimum. The sum of the two expected regrets is 2κ times that testing-error sum. Their maximum is at least their average, which proves the first inequality. Independent observations having identical laws in the two worlds do not change the total variation distance. For the second inequality, the KL divergence of one annotated pair is

$$\mathrm{KL}(D_+ \| D_-) = \tfrac{1}{2}\{\mathrm{kl}(\pi + a, \pi - a) + \mathrm{kl}(\pi - a, \pi + a)\} = K,$$

where $\mathrm{kl}(u,v) = u\log(u/v) + (1-u)\log[(1-u)/(1-v)]$ is Bernoulli KL. Independence gives product KL nK. For completeness, the finite-space log-sum inequality reduces KL to the Bernoulli KL of an event attaining total variation. For fixed v, the second derivative of $\mathrm{kl}(u,v)$ with respect to u is $1/[u(1-u)] \geq 4$. Its value and first derivative at $u = v$ are zero, so $\mathrm{kl}(u,v) \geq 2(u-v)^2$. Hence TV is at most $\sqrt{nK/2}$. The constant score has AUC $1/2$ in both worlds, proving attainment when $n = 0$.

Adaptive annotation. The bound extends to active label acquisition. At each of n steps the auditor may, as a function of all earlier observations and independent randomness, reveal the label of a chosen item, reveal the label of a fresh pool draw, or receive a fresh draw from the error class or

from the correct class. Let H_n^σ be the law of the complete history. The query rule is the same in both worlds, so by the chain rule for KL divergence $\mathrm{KL}(H_n^+ \| H_n^-) \le nK_{\mathrm{ad}}$, where

$$K_{\mathrm{ad}} = \max\left\{\mathrm{kl}(\pi+a,\pi-a),\ \mathrm{kl}(\pi-a,\pi+a),\ \frac{a}{\pi}\log\frac{\pi+a}{\pi-a},\ \frac{a}{1-\pi}\log\frac{1-\pi+a}{1-\pi-a}\right\}$$

bounds the conditional divergence of any single observation: the labels of the two items and the item drawn from each class. A fresh pool label contributes their average K. Replacing D_σ^n by H_n^σ in the testing argument gives

$$\max_\sigma \mathbb{E}_{H_n^\sigma}\mathcal{R}_\sigma(\widehat{s}) \ge \kappa\{1-\mathrm{TV}(H_n^+,H_n^-)\} \ge \kappa\max\{0,1-\sqrt{nK_{\mathrm{ad}}/2}\}.$$

Adaptive and class-conditional queries can separate the worlds faster than i.i.d. labels, but no query rule exceeds this information budget. An exact dynamic program over every deterministic adaptive policy with $n \le 6$, at four settings of (π, a), confirms the bound and shows the gain over i.i.d. annotation.

This is an explicit audit-regret specialization of classical two-point testing lower bounds (Tsybakov, 2009), not a new testing inequality. It supplies a nonzero separation and a finite-label dependence to the identification witness, under i.i.d. or adaptive annotation, and is not a numerical sample-size certificate for EMMA.

E.2 Cell-only AUC and the support witness

Results 2 and 3: ties and balance. For a cell-only score, put $h_{ab} = \mathbf{1}[s_a > s_b] + \frac{1}{2}\mathbf{1}[s_a = s_b]$. Conditioning on both cells gives

$$\mathrm{AUC}(s) = \sum_{a,b} e_a c_b h_{ab}, \qquad h_{ab} + h_{ba} = 1. \tag{13}$$

This includes $h_{aa} = 1/2$. If $e = c$, exchanging a and b and adding the two expressions yields $2\mathrm{AUC} = \sum_{a,b} e_a e_b = 1$.

Result 4: exact maximum and efficient computation. Using $D_{ab} = e_a c_b - e_b c_a$,

$$\mathrm{AUC}(s) - \tfrac{1}{2} = \tfrac{1}{2}\sum_{a<b} D_{ab}(h_{ab} - h_{ba}) \le \tfrac{1}{2}\sum_{a<b}|D_{ab}| = \Lambda.$$

Ordering by e_a/c_a makes every unequal pair contribute positively and attains the bound. For zero denominators, use the equivalent score $e_a/(e_a + c_a)$. Cells with both masses zero are irrelevant. Reversing the order attains $1/2 - \Lambda$. Consequently $0 \le \Lambda \le 1/2$ and $\Lambda = 0$ if and only if $e = c$. The converse follows by summing $D_{ab} = 0$ over b, which gives $e_a - c_a = 0$. With $\mathrm{TV}(P,Q) = \frac{1}{2}\sum_i |P_i - Q_i|$,

$$\Lambda = \tfrac{1}{2}\mathrm{TV}(e \otimes c, c \otimes e), \qquad (e \otimes c)_{ab} = e_a c_b.$$

This is the categorical maximum of Chen et al. (2021b, Observation 3).

Computation. Sort cells in ascending order of $e_a/(e_a+c_a)$ and define prefix masses $E_{<j} = \sum_{i<j} e_i$ and $C_{<j} = \sum_{i<j} c_i$. Then

$$\Lambda = \tfrac{1}{2}\sum_{j=1}^{k}(e_j C_{<j} - c_j E_{<j}). \tag{14}$$

Sorting costs $O(k \log k)$ and the prefix pass costs $O(k)$.

Result 4a: refinement. If $T = f(T')$ for some function f, every score based on T is also based on T'. Taking the maximum AUC over these nested score classes proves $\Lambda(T') \ge \Lambda(T)$. A bound on a fine map bounds its coarsenings. A bound on bins of a continuous score bounds only scores constant on those bins unless the partition preserves the original score exactly.

Result 4b: stability and a finite-sample population certificate. For probability vectors e, c, e', c' on the same k cells,

$$|\Lambda(e,c) - \Lambda(e',c')| \le \mathrm{TV}(e,e') + \mathrm{TV}(c,c'). \tag{15}$$

Indeed, the reverse triangle inequality applied to $2\Lambda(e,c) = \mathrm{TV}(e \otimes c, c \otimes e)$ bounds its change by $\mathrm{TV}(e \otimes c, e' \otimes c') + \mathrm{TV}(c \otimes e, c' \otimes e')$. Each product distance is at most $\mathrm{TV}(e,e') + \mathrm{TV}(c,c')$, by inserting an intermediate product and using $\mathrm{TV}(u \otimes v, u' \otimes v) = \mathrm{TV}(u,u')$. Dividing by two proves the claim, including boundary cells with zero mass.

Finite-sample certificate. Now fix the cell map before observing the evaluation sample. Suppose each class supplies $n_y \ge 1$ independent, identically distributed, unweighted observations, with cell laws e for $y = 1$ and c for $y = 0$. Let $\widehat{e}, \widehat{c}$ be relative counts, $0 < \beta < 1$ the desired failure probability, and define

$$b(k,n,\beta) = \min\left\{1, \sqrt{\frac{k\log 2 + \log(2/\beta)}{2n}}\right\}, \tag{16}$$

$$r(k,n_1,n_0,\beta) = \min\{\tfrac{1}{2}, b(k,n_1,\beta) + b(k,n_0,\beta)\}. \tag{17}$$

Then, with $\widehat{\Lambda} = \Lambda(\widehat{e},\widehat{c})$,

$$\Pr\{|\widehat{\Lambda} - \Lambda(e,c)| \le r(k,n_1,n_0,\beta)\} \ge 1 - \beta. \tag{18}$$

The confidence set is $[\widehat{\Lambda} - r, \widehat{\Lambda} + r] \cap [0, 1/2]$.

Concentration proof. For any class law u on k cells, $\mathrm{TV}(\widehat{u}, u) = \max_{A \subseteq \{1,\dots,k\}} [\widehat{u}(A) - u(A)]$. For a fixed subset A, the empirical mass is an average of independent indicators. The bounded-variable exponential moment inequality $\mathbb{E}e^{v(B - \mathbb{E}B)} \le e^{v^2/8}$ for $0 \le B \le 1$, followed by exponential Markov and minimization in v, gives $\Pr(\widehat{u}(A) - u(A) > s) \le e^{-2ns^2}$. A union bound over 2^k subsets gives $2^k e^{-2ns^2}$. Allocate failure probability $\beta/2$ to each class, then apply (15). Clipping uses $\mathrm{TV} \le 1$ and $|\widehat{\Lambda} - \Lambda| \le 1/2$. For $k = 1$, both diagnostics are identically zero and radius zero is available.

Scope. This certificate can be vacuous for many cells or small classes. It applies to the plug-in statistic, not the cross-fit estimator. Random class counts may be conditioned on when sampling is i.i.d. from the mixture. A map learned on independent training data may likewise be conditioned on. Nominal calibration weights, effective sample counts, document dependence and maps chosen using the same evaluation labels do not satisfy this statement without additional arguments.

Result 3a: an empirical support floor. This bounds Λ from below, so it is stated after Result 4, which defines Λ. Let $u_e = \sum_{a:c_a=0} e_a$ and $u_c = \sum_{a:e_a=0} c_a$. The score $s_a = \mathbf{1}[c_a = 0]$ ranks error-only cells above every correct item, while all remaining comparisons tie. Its AUC is $u_e + (1 - u_e)/2 = 1/2 + u_e/2$. The score $-\mathbf{1}[e_a = 0]$ supplies the other witness. Hence

$$\Lambda(e,c) \ge \tfrac{1}{2}\max(u_e, u_c). \tag{19}$$

Scope. Applied to empirical counts, this statement concerns the empirical distributions. Sampling zeros need not be population zeros. On the CAGED FB15k-237 signature map, $u = 0.923$ yields an indicator AUC of 0.9615, beside the type prior's 0.954 measured at its own resolution.

Selected sets. For a finite candidate pool, let A candidates lie on signatures attested in the correct class and require a selected set of N distinct anomalies. Any such set contains at least $\max(0, N - A)$ unattested candidates, giving

$$\Lambda_{\text{selected}} \ge \tfrac{1}{2}\max(0, 1 - A/N).$$

This is a restriction on that candidate pool and required count. It does not rule out generating more candidates. Here 7.7% are attested, so replenishing attested candidates costs approximately thirteen times as many draws. That estimate alone does not guarantee matching every cell marginal.

E.3 FIDELITY AND RANKING REGRET

Result 5: global bound. Assume p_{err} and q_θ are strictly positive P-almost surely, so the following scores and their differences are finite on both audit classes. Let $d^* = \log(p_{\mathrm{err}}/p)$, $d_\theta = \log(q_\theta/p)$ and $\delta = d_\theta - d^* = \log(q_\theta/p_{\mathrm{err}})$, with $\varepsilon = \sup_x |\delta(x)| < \infty$. Then

$$0 \le \mathcal{R}_\theta := \mathrm{AUC}(d^*) - \mathrm{AUC}(d_\theta) \le \Pr(0 \le M \le 2\varepsilon), \qquad M = d^*(X^+) - d^*(X^-). \tag{20}$$

Here and below X^+ and X^- are independent draws from the error and correct classes. *Proof.* The lower bound is likelihood-ratio optimality. For the upper bound, $|\delta(X^+) - \delta(X^-)| \le 2\varepsilon$, so a strictly positive oracle margin can lose comparison credit only inside the stated band. A negative oracle margin cannot lose credit. An oracle tie is included by the closed lower endpoint. Each pair loses at most one unit of credit. Integrating proves the bound. At $\varepsilon = 0$ the two scores coincide and regret is zero, although the displayed band may contain oracle ties.

Result 5a: local bound. For any measurable set B, set $\varepsilon_B = \sup_{x \in B} |\delta(x)|$. Charging pairs with an endpoint outside B in full and using the previous argument inside it gives

$$\mathcal{R}_\theta \le P_{\mathrm{err}}(B^c) + P_{\mathrm{pos}}(B^c) + \Pr(X^+, X^- \in B,\ 0 \le M \le 2\varepsilon_B). \tag{21}$$

The first two terms use a union bound. Every choice of B is valid, so an infimum over candidate sets remains a population bound. Replacing unknown masses and margins by estimates does not itself provide a confidence guarantee.

Result 5b: removing guaranteed within-cell ties. Assume *both* d^* and d_θ are constant on every cell of T. All diagonal pairs then receive half credit under both scores, while the band in (20) counts them. Therefore

$$\mathcal{R}_\theta \le \Pr(0 \le M \le 2\varepsilon) - \sum_a e_a c_a. \tag{22}$$

For the local version, if B is a union of cells, subtract only $\sum_{a:\, T^{-1}(a) \subseteq B} e_a c_a$ from (21). A cell-constant *tilt* is insufficient: take one coarse cell containing two items, $p_{\mathrm{pos}} = (0.8, 0.2)$, $p_{\mathrm{err}} = (0.2, 0.8)$, $\pi = 1/2$, and $q_\theta = q_0 = (0.8, 0.2)$. The correction is constant, but the oracle is not. Regret is 0.60, the global band is 0.96, and subtracting the coarse diagonal 1 gives the invalid value -0.04. Singleton cells have diagonal 0.32 and restore validity. Our measured bound concerns the cell-only surrogate where the hypothesis holds. In a 40000-instance search, the largest ratio of regret to the valid tightened bound is 99.7%, versus 90.2% for the original band.

Result 5c: average perturbation and pair margins. Retain the independent audit draws and finite scores of Result 5, but do not assume a finite supremum error. Define $\Delta = \delta(X^+) - \delta(X^-)$ and $M = d^*(X^+) - d^*(X^-)$. For every threshold $\tau > 0$,

$$0 \le \mathcal{R}_\theta \le \Pr(0 \le M \le \tau) + \Pr(|\Delta| > \tau) \tag{23}$$

$$\le \Pr(0 \le M \le \tau) + \frac{\mathbb{E}_E|\delta(X)| + \mathbb{E}_C|\delta(X)|}{\tau}. \tag{24}$$

Here $\mathbb{E}_E$ and $\mathbb{E}_C$ denote expectations under the error and correct-item laws. The second bound is useful when both first moments are finite. If both second moments are finite, an alternative tail term is

$$\Pr(|\Delta| > \tau) \le \frac{\mathbb{E}\Delta^2}{\tau^2}, \qquad \mathbb{E}\Delta^2 = \mathrm{Var}_E(\delta) + \mathrm{Var}_C(\delta) + (\mathbb{E}_E\delta - \mathbb{E}_C\delta)^2.$$

The variance identity uses independence of the two class draws. It is invariant to adding a common constant to the perturbation, unlike the looser sum of absolute moments.

Proof. The fitted margin is $M + \Delta$. If $M > \tau$ and $|\Delta| \le \tau$, the oracle and fitted comparisons both earn full credit. If $M < 0$, oracle credit is zero and cannot decrease. All possible positive credit loss is therefore contained in $\{0 \le M \le \tau\} \cup \{|\Delta| > \tau\}$, and each pair loses at most one. Integrate and apply the first- or second-moment Markov inequality to the tail. The closed margin band includes oracle ties, so the bound remains valid for discrete distributions. Each upper bound may be clipped at one and minimized over $\tau > 0$. These are population inequalities involving unknown target margins and perturbations. No new empirical certificate is claimed. When $|\delta| \le \varepsilon$, taking $\tau = 2\varepsilon > 0$ recovers Result 5. At $\varepsilon = 0$, the scores coincide and regret is zero directly.

Result 5d: total-variation fidelity and contrastive fitting error. Fix $P = \pi E + (1-\pi)C$ with $0 < \pi < 1$, and any auxiliary probability law $Q \ll P$. The latter condition means Q assigns no mass to P-null sets. It does not require Q to be a possible mixture component at prevalence π. Let

$$A(s, U, V) = \mathbb{E}_{X\sim U,\, X'\sim V}\left[\mathbf{1}[s(X) > s(X')] + \tfrac{1}{2}\mathbf{1}[s(X) = s(X')]\right]$$

for independent draws, with U positive and V negative. Scores are measurable, real-valued and otherwise unrestricted. Set $A^*(U, V) = \sup_v A(v, U, V)$ and define

$$d_{\mathrm{TV}} = \mathrm{TV}(E, Q), \qquad \epsilon_{\mathrm{ctr}}(s) = A^*(Q, P) - A(s, Q, P) \geq 0.$$

Then the true audit regret of any such score satisfies

$$0 \leq \mathcal{R}(s) := A^*(E, C) - A(s, E, C) \leq \min\left\{1, \frac{2d_{\mathrm{TV}} + \epsilon_{\mathrm{ctr}}(s)}{1-\pi}\right\}. \tag{25}$$

In particular, $s_Q = dQ/dP$ attains $A^*(Q, P)$ and has zero contrastive shortfall, even when Q has zeros. An integrable density ratio is finite P-almost surely and may be assigned any finite value on the remaining null set. Using the ratio itself avoids requiring finite log scores.

Proof. Symmetry with half credit for ties gives $A(s, E, E) = 1/2$. By the mixture identity, for every score,

$$A(s, E, P) = (1-\pi)A(s, E, C) + \pi/2. \tag{26}$$

Taking suprema preserves this affine identity. For fixed s, define

$$g_s(x) = \mathbb{E}_{X'\sim P}\left[\mathbf{1}[s(x) > s(X')] + \tfrac{1}{2}\mathbf{1}[s(x) = s(X')]\right].$$

Since $0 \leq g_s \leq 1$, the variational bound for total variation gives $|A(s, E, P) - A(s, Q, P)| = |\mathbb{E}_E g_s - \mathbb{E}_Q g_s| \leq d_{\mathrm{TV}}$. The same bound holds between the two suprema. Consequently

$$\begin{aligned}(1-\pi)\mathcal{R}(s) &= A^*(E, P) - A(s, E, P)\\ &= [A^*(E, P) - A^*(Q, P)] + \epsilon_{\mathrm{ctr}}(s) + [A(s, Q, P) - A(s, E, P)]\\ &\leq 2d_{\mathrm{TV}} + \epsilon_{\mathrm{ctr}}(s).\end{aligned}$$

The range of AUC supplies clipping at one. Likelihood-ratio optimality for the binary problem (Q, P) proves the statement about s_Q.

Interpretation and scope. Closeness to the true error law and successful contrastive ranking jointly control audit regret. The shortfall is relative to the unrestricted population optimum, so it includes representation or model-class limitations at the stated resolution as well as fitting error. Neither empirical training AUC nor logistic training loss is that shortfall. Apply all distributions at a common fixed representation if the score retains only that information. The bound is conditional on a realized fitted score and does not estimate its unknown terms, certify the deployed detectors, or require $P - \pi Q$ to be nonnegative. Its proof is a bounded-loss transfer argument and the mixture AUC identity, not a claim that total-variation stability is new. Small TV is sufficient, not necessary: substantially different auxiliaries can preserve the same audit ordering. The proof uses E and Q only through the bounded comparison functions g_v of measurable scores v. Hence $\mathrm{TV}(E, Q)$ may be replaced throughout by $d_{\mathcal{G}}(E, Q) = \sup_v |\mathbb{E}_E g_v - \mathbb{E}_Q g_v| = \sup_v |A(v, E, P) - A(v, Q, P)|$, an integral probability metric over this class with $d_{\mathcal{G}} \leq \mathrm{TV}$. The fixed-score step is bounded by $d_{\mathcal{G}}$ directly, and the difference of the two suprema by the supremum of the differences. This is the form stated as Theorem 2. The refinement is at most a factor of two: for $v = \mathbf{1}_B$, $g_v = 1 - P(B)/2$ on B and $(1 - P(B))/2$ off it, so $\mathbb{E}_E g_v - \mathbb{E}_Q g_v = \frac{1}{2}[E(B) - Q(B)]$ and $d_{\mathcal{G}} \geq \frac{1}{2}\mathrm{TV}(E, Q)$.

Result 5e: conditional fidelity under matched stratum marginals. Let G be a finite observed stratum, with $p_g = P(G = g) > 0$ and $0 < \pi_g = \Pr(Y = 1 \mid G = g) < 1$. Write P_g, E_g, C_g for the conditional item laws within stratum g, so $P_g = \pi_g E_g + (1-\pi_g)C_g$. Let $Q_g \ll P_g$ be an auxiliary law within that stratum, and define joint laws $P(g, dx) = p_g P_g(dx)$ and $Q(g, dx) = p_g Q_g(dx)$ on the disjoint union of strata. For any measurable real score $s(g, x)$ and fixed weights $\omega_g \geq 0$ summing to one, define conditional audit regret and contrastive shortfall by

$$\mathcal{R}_\omega(s) = \sum_g \omega_g\{A^*(E_g, C_g) - A(s_g, E_g, C_g)\}, \qquad \epsilon_g(s) = A^*(Q_g, P_g) - A(s_g, Q_g, P_g),$$

where $s_g(x) = s(g, x)$ and A gives half credit to ties. Then

$$\mathcal{R}_\omega(s) \le \sum_g \omega_g \min\left\{1, \frac{2\mathrm{TV}(E_g, Q_g) + \epsilon_g(s)}{1 - \pi_g}\right\}. \tag{27}$$

Moreover, $dQ/dP(g, x) = dQ_g/dP_g(x)$. If $Q_g = E_g$ for every stratum, this ratio is $\eta_g(x)/\pi_g$, where $\eta_g(x) = \Pr(Y = 1 \mid G = g, X = x)$, and is optimal for every within-stratum audit simultaneously. This need not be the pooled oracle ordering because the factors $1/\pi_g$ differ across strata.

Proof. Apply Result 5d to each conditional mixture and sum with the nonnegative weights. Matching joint stratum masses cancels p_g in the density ratio. Bayes' rule gives $dE_g/dP_g = \eta_g/\pi_g$, proving the ordering claim. Since strata form disjoint measurable sets, their oracle scores can be combined into one joint score, so the sum of their optimal AUCs is jointly attainable. For AUC conditioned on an independent error/correct pair sharing a stratum,

$$\omega_g = \frac{E(G = g)C(G = g)}{\sum_h E(G = h)C(G = h)} = \frac{p_g^2 \pi_g(1 - \pi_g)}{\sum_h p_h^2 \pi_h(1 - \pi_h)}.$$

This follows by conditioning the independent pair's group probabilities. The QA experiment uses the empirical analogue, not uniform weights over generators.

Representation and estimation boundary. The oracle statement allows the score to use the stratum. A fitted feature-only head can discard that information or be misspecified, and its ϵ_g need not vanish. The inequality still applies to that head. Neither matching group masses nor high training AUC measures the unknown conditional fidelity terms. This is a corollary for a specified conditional target, not a guarantee of unconditional improvement.

Balanced-permutation control. In a finite fitting sample of n rows, let stratum g have n_g rows and $e_g > 0$ errors. Assign each error auxiliary mass $(n_g/n)/e_g$, and uniformly permute the labels within each stratum, preserving e_g. Row i in stratum g is then selected with probability e_g/n_g. Its expected auxiliary mass is exactly $(e_g/n_g)(n_g/n)/e_g = 1/n$. Consequently $\mathbb{E}_{\mathrm{perm}}\widehat{Q}_{\mathrm{bal,perm}} = \widehat{P}$ as a row-indexed empirical law. Applying any fixed feature map preserves this equality. Unlike unbalanced stratified permutation, this control removes stratum-rate information in expectation. A single permutation may deviate from that expectation, and fitting is nonlinear, so no equality of fitted-score expectations is asserted.

Result 5f: transferring an audit across a change of pool. Let $P = \pi E + (1 - \pi)C$ be the target pool, $0 < \pi < 1$, and let P_0 be the training-pool law on the same measurable representation. For a training auxiliary $Q \ll P_0$ and real measurable score s, define $\epsilon_0(s) = A^*(Q, P_0) - A(s, Q, P_0)$. Then

$$\mathcal{R}_{\mathrm{target}}(s) \le \min\left\{1, \frac{2\mathrm{TV}(E, Q) + 2\mathrm{TV}(P, P_0) + \epsilon_0(s)}{1 - \pi}\right\}. \tag{28}$$

Proof. The bounded half-tie comparison kernel gives, for every score v,

$$|A(v, E, P) - A(v, Q, P_0)| \le \mathrm{TV}(E, Q) + \mathrm{TV}(P, P_0).$$

Insert $A(v, Q, P)$ and apply the variational TV bound once to each independent argument. The bound also holds for the suprema over v. Substituting it into the three-term decomposition in Result 5d and using the target mixture identity proves (28). Applying the argument within each stratum and summing proves the corresponding shifted-pool version of Result 5e.

The pool term is necessary. Take $\pi = 1/2$, $E = Q = (0.8, 0.2)$, $C = (0.2, 0.8)$, $P = (0.5, 0.5)$ and $P_0 = (0.9, 0.1)$. The optimal training score $Q/P_0 = (8/9, 2)$ reverses the target ordering, with target AUC 0.2 against oracle 0.8, despite perfect error-law fidelity and zero training contrastive shortfall. Thus a cross-task transfer claim cannot silently identify its training pool with its target pool. The inequality is a population transfer statement with unknown discrepancies, not a certificate obtained by plugging in empirical distributions or training loss.

E.4 Fitting the tilt and characterizing optimal orderings

Result 6: fitting objective, derivatives and score identity. Let $(x_i, w_i)_{i=1}^m$ be annotated errors with nonnegative weights and positive total weight, and let $(z_j, v_j)_{j=1}^M$ represent a corruption sample with $v_j \geq 0$ and $\sum_j v_j = 1$. For a ridge parameter $\lambda \geq 0$, define

$$\overline{\varphi}_{\mathrm{err}} = \frac{\sum_i w_i \varphi(x_i)}{\sum_i w_i}, \qquad \widehat{Z}(\theta) = \sum_{j=1}^{M} v_j e^{\theta^\top \varphi(z_j)}, \qquad J(\theta) = \log \widehat{Z}(\theta) - \theta^\top \overline{\varphi}_{\mathrm{err}} + \frac{\lambda}{2}\|\theta\|^2.$$

Derivatives. With tilted sample probabilities $\widetilde{v}_j = v_j e^{\theta^\top \varphi(z_j)} / \widehat{Z}(\theta)$,

$$\nabla J(\theta) = \sum_j \widetilde{v}_j \varphi(z_j) - \overline{\varphi}_{\mathrm{err}} + \lambda\theta,$$
$$\nabla^2 J(\theta) = \mathrm{Cov}_{\widetilde{v}}(\varphi) + \lambda I_d.$$

Here I_d is the d-dimensional identity matrix and the covariance is the weighted covariance of the feature vectors. The Hessian is positive semidefinite, and positive definite when $\lambda > 0$. Thus a finite stationary point satisfies the penalized moment equation. For $\lambda = 0$ it matches moments exactly, when such a finite optimizer exists. The finite-sample derivatives above use finite feature vectors and nonnegative normalized corruption weights. For population derivatives, take θ in the interior of the natural parameter domain $\{\vartheta : Z(\vartheta) < \infty\}$. Finiteness in a neighborhood justifies the first two derivatives of the log normalizer and their moment/covariance formulas. A finite normalizer at a boundary point alone is not such a justification. Under this condition, replace sample expectations by expectations under q_θ. The ridge penalty absorbs 3.5% of the log-rate spread in our fit, below the prespecified 15% cap.

Cell features. For one-hot cell features, an unpenalized fit with positive base and error mass in every cell has $\theta_a = \log(\widehat{e}_a / q_{0,a}) + C$ for an arbitrary common constant C. This explains why a coarse tilt may induce the same ordering as a simple prior. It is not a general identity with the conditional error rate of each relation. Finally,

$$\log \frac{q_\theta(x)}{q_0(x)} = \theta^\top \varphi(x) - \log Z(\theta)$$

proves (33). Only the constant can be omitted without changing rankings. An arbitrary transformation of the base score does not preserve this additive identity.

Results 7 and 7a: sufficient class and exact finite-cell boundary. For a positive strictly increasing function h with finite normalizer, choosing $q \propto p_{\mathrm{pos}} h(\rho)$ makes q/p_{pos} a strictly increasing function of ρ and therefore AUC-optimal. This statement contains no prevalence parameter when the two class-conditional distributions are fixed. It does not imply the same result for q/p (Result 1a).

The full discrete characterization is the classical likelihood-ratio result (Menon & Williamson, 2016, Proposition 42), restated in pair masses. With $w_{ab} = e_a c_b$,

$$s \text{ is AUC-optimal} \iff \mathrm{sign}(s_a - s_b) = \mathrm{sign}(w_{ab} - w_{ba}) \quad \text{whenever } w_{ab} \neq w_{ba}. \tag{29}$$

Proof. Each unordered pair contributes its larger weight when ordered correctly, its smaller weight when reversed, and their average when tied. Maximizing every unequal pair proves both directions. Equal-weight pairs impose no restriction. For example, $e = (0.5, 0.5, 0)$ and $c = (0.25, 0.25, 0.5)$ have likelihood ratios $(2, 2, 0)$. Both scores $(2, 2, 0)$ and $(2, 1, 0.5)$ attain AUC 0.75, although the latter separates equal-ratio cells and is not a function of the ratio alone. The contribution here is the corruption-design interpretation and checkability on marginals, not a new characterization.

Budget illustration. In the simulation of Appendix AA, 100 errors recover the oracle ordering in 93.2% of runs and 200 in 99.0%. At those budgets, the 95th percentile log-density errors remain 0.558 and 0.344. Reaching $\varepsilon < 0.25$ at that percentile takes 400 errors. This illustrates the distinction between recovering an ordering and accurately estimating every density, under that simulation's assumptions.

F ESTIMATORS, UNCERTAINTY AND REPORTING CONVENTIONS

Weighted counts and AUC. For labelled rows (x_i, y_i, w_i) with finite nonnegative weights, define $W_y = \sum_{i:y_i=y} w_i > 0$ and $n_y = \#\{i : y_i = y\}$. Unresolved labels are excluded, not counted as correct. The empirical marginals and weighted AUC are

$$\widehat{e}_a = \frac{\sum_{i:y_i=1} w_i \mathbf{1}[T(x_i) = a]}{W_1}, \qquad \widehat{c}_a = \frac{\sum_{i:y_i=0} w_i \mathbf{1}[T(x_i) = a]}{W_0}.$$

$$\widehat{\mathrm{AUC}}_w(s) = \frac{\sum_{i:y_i=1} \sum_{j:y_j=0} w_i w_j \big(\mathbf{1}[s_i > s_j] + \frac{1}{2}\mathbf{1}[s_i = s_j]\big)}{W_1 W_0}.$$

Setting all weights to one recovers the ordinary empirical quantities. For our enriched sheets the weights are nominal calibration weights. They do not reconstruct a known probability sampling design. The plug-in diagnostic is $\widehat{\Lambda}_{\mathrm{plug}} = \Lambda(\widehat{e}, \widehat{c})$.

Cross-fitting. For each repetition $r = 1, \ldots, R$, split the observations within each class into two halves. Normalize class weights within each half to obtain $\widehat{e}^{(r,j)}, \widehat{c}^{(r,j)}$, $j \in \{1, 2\}$, and form

$$\widehat{D}^{(r,j)}_{ab} = \widehat{e}^{(r,j)}_a \widehat{c}^{(r,j)}_b - \widehat{e}^{(r,j)}_b \widehat{c}^{(r,j)}_a.$$

The reported cross-fitted sensitivity is

$$\widehat{\Lambda}_{\mathrm{cf}} = \frac{1}{4R} \sum_{r=1}^{R} \sum_{a<b} \Big[\mathrm{sign}(\widehat{D}^{(r,1)}_{ab}) \widehat{D}^{(r,2)}_{ab} + \mathrm{sign}(\widehat{D}^{(r,2)}_{ab}) \widehat{D}^{(r,1)}_{ab} \Big], \tag{30}$$

with $\mathrm{sign}(0) = 0$ and $R = 20$ by default. A sign learned on one half is evaluated on the other, reducing the plug-in's reuse of noise to choose and measure an ordering. Incorrect signs attenuate contributions. Repeating splits reduces splitting variability, not this attenuation. The estimator can be negative near a null and is not an unbiased estimator or a generally valid lower bound.

Interpretation of the cross-fitted value. Call $\widehat{\Lambda}_{\mathrm{cf}}$ a *cross-fitted pairwise sensitivity statistic*. (30) is well defined for splits carrying positive total weight in each class in each half, even when individual cells are absent from a fitting half. A half holding no member of a class has no normalised marginal at all, which is a stronger failure than the cell-level condition below. It cannot arise for any row we report, since every reported row has at least two members of each class.

What is conditional is interpreting the statistic as the performance of one scalar score. Fix a repetition and a direction, suppressing the index r, and write $\Gamma_{j\to k} = \frac{1}{2} \sum_{a<b} \mathrm{sign}(\widehat{D}^{(j)}_{ab}) \widehat{D}^{(k)}_{ab}$ for fitting on half j and evaluating on half k, so that $\widehat{\Lambda}_{\mathrm{cf}}$ is the average of $\Gamma_{j\to k}$ over both directions of all R repetitions. Result 4's derivation already puts any cell-only score's AUC in pairwise form, and since $h_{ab} - h_{ba} = \mathrm{sign}(s_a - s_b)$ it takes the form

$$\mathrm{AUC}(s) - \tfrac{1}{2} = \tfrac{1}{2} \sum_{a<b} \mathrm{sign}(s_a - s_b)\, D_{ab}.$$

The fitted cell ratio $t^{(j)}_a = \widehat{e}^{(j)}_a / (\widehat{e}^{(j)}_a + \widehat{c}^{(j)}_a)$ satisfies

$$t^{(j)}_a - t^{(j)}_b = \frac{\widehat{D}^{(j)}_{ab}}{(\widehat{e}^{(j)}_a + \widehat{c}^{(j)}_a)(\widehat{e}^{(j)}_b + \widehat{c}^{(j)}_b)},$$

so $\mathrm{sign}(t^{(j)}_a - t^{(j)}_b) = \mathrm{sign}(\widehat{D}^{(j)}_{ab})$ whenever both denominators are positive. A cell holding only errors or only correct items in the fitting half is not a problem: it scores 1 or 0. The requirement is

$$\widehat{e}^{(k)}_a + \widehat{c}^{(k)}_a > 0 \quad \Longrightarrow \quad \widehat{e}^{(j)}_a + \widehat{c}^{(j)}_a > 0, \tag{31}$$

for every cell, in every direction. We call this *fitting coverage*. It is a property of the splits and is unrelated to the two other senses this paper gives "coverage": the support condition $q > 0$ P-almost surely of §3, and the interval coverage of Table 22. Under it $\frac{1}{2} + \Gamma_{j\to k}$ is exactly the held-out AUC of the fitting half's ratio order, and if it holds in both directions of all R repetitions then $\frac{1}{2} + \widehat{\Lambda}_{\mathrm{cf}}$ is

the *mean of the directional held-out AUCs*. That mean is not the AUC obtained by pooling every out-of-fold prediction, because the directions rank by differently fitted scores.

Three qualifications matter. **Coverage is sufficient but not necessary.** When a cell carries evaluation mass that is absent from the fitting half, $\widehat{D}_{ab}^{(j)} = 0$ against every partner, so those pairs abstain, and the abstentions need not be jointly realizable by any scalar score. Take three cells, a fitting half holding error counts $(10, 0, 0)$ against correct counts $(0, 10, 0)$, and an evaluation half holding $(4, 2, 4)$ against $(2, 5, 3)$. The third cell carries evaluation mass and no fitting mass, and the direction implies 0.58, which *no* score on three cells attains. **The empirical floor does not detect this case.** Pooling those same counts to $(14, 2, 4)$ against $(2, 15, 3)$ gives floor 0 and plug-in 0.3675, so (19) and (31) are separate conditions and neither is a population guarantee. **The identity applies to the raw statistic.** Where a negative estimate is reported as zero, the interpretation does not survive the truncation: two fully covered halves with opposite class profiles have mean held-out AUC 0.20 at a raw statistic of -0.30, which truncation would present as 0.50. Every row we report therefore carries the raw value, the plug-in, a flag recording whether any truncation was applied, the covered and total direction counts, and the held-out-AUC interpretation only where it is established. Where it is not, the pairwise statistic stands alone and $\frac{1}{2} + \widehat{\Lambda}_{\mathrm{plug}}$ remains the exact maximum for the supplied rows. Removing offending folds or cells to make (31) pass would change the statistic, and we do not do it. The numerical checks verify that identity, the condition, both counterexamples and the agreement of the production estimator with an independent held-out AUC calculation. The reproduction gate runs it.

Analytic sensitivity scale. Away from $D_{ab} = 0$, let $S_{ab} = \mathrm{sign}(D_{ab})$, with zero diagonal. The gradients of the population functional are

$$g_e = \tfrac{1}{2} S c, \qquad g_c = -\tfrac{1}{2} S e.$$

For independent multinomial samples of sizes n_1 and n_0, the first-order variance is

$$V = \frac{\sum_a e_a g_{e,a}^2 - (\sum_a e_a g_{e,a})^2}{n_1} + \frac{\sum_a c_a g_{c,a}^2 - (\sum_a c_a g_{c,a})^2}{n_0}. \tag{32}$$

We plug in empirical marginals and use the interval

$$\left[\widehat{\Lambda}_{\mathrm{cf}} - z\sqrt{\widehat{V}},\ \widehat{\Lambda}_{\mathrm{cf}} + z\sqrt{\widehat{V}}\right],$$

without clipping its endpoints. The default multiplier is $z = 3$, not the usual normal 95% multiplier. At zero pair differences the derivative is non-regular. Moreover the scale is centered on a potentially attenuated cross-fit estimate. Table 22 measures limited simulation coverage and includes undercoverage. This construction is an analytic sensitivity interval, not a general 95% confidence procedure. For weighted data, the implementation replaces n_y by

$$n_{\mathrm{eff},y} = \frac{W_y^2}{\sum_{i:y_i=y} w_i^2}.$$

This is a Kish approximation and does not repair unknown inclusion probabilities or within-cluster dependence.

Support reporting and permutation nulls. The empirical floor is $F = \frac{1}{2}\max(\widehat{u}_e, \widehat{u}_c)$ from (19). If the cross-fitted value is undefined or below a positive floor, the point summary is declined. Its reported bracket $[F, \widehat{\Lambda}_{\mathrm{plug}}]$ compares two empirical summaries. It is not a confidence set for a population value. When the floor is zero, a negative cross-fit value may instead be reported as zero if the minority observations per cell are at least 20, the simulated density threshold. The density flag is reported separately. Neither clearing this diagnostic nor meeting that threshold certifies accurate estimation. For the benchmark survey, labels are permuted and the same feature-family and binary-contrast selection is repeated on each permutation. The 95th percentile of these null summaries is compared with the observed summary. This is a conditional diagnostic for the specified search, not a bound on all partial-input models.

Prespecified map families and selection-adjusted permutations. Fix a finite family $\mathcal{T} = \{T_1, \ldots, T_m\}$ before inspecting evaluation labels, with $m \geq 1$, and report every map and its cell count. The population quantity and its empirical statistic are, respectively,

$$\Lambda_{\max} = \max_{1 \leq j \leq m} \Lambda(T_j), \qquad S_{\text{plug}} = \max_{1 \leq j \leq m} \widehat{\Lambda}_{\text{plug}}(T_j), \qquad S_{\text{cf}} = \max_{1 \leq j \leq m} \widehat{\Lambda}_{\text{cf}}(T_j).$$

The two empirical maxima are different statistics and are not interchangeable below. Our survey prespecifies the cross-fitted one, so $S_{\text{obs}} = S_{\text{cf}}$ there. The permutation argument needs only that the choice is fixed in advance and recomputed identically inside every permutation. Under the null that labels are exchangeable conditional on the observed features and class counts, permute the labels and recompute all maps' statistics and their maximum on each permutation. For B_{perm} independent uniform label permutations, let S_b be the recomputed maximum on permutation b. The Monte Carlo permutation p-value is

$$p_{\text{perm}} = \frac{1 + \sum_{b=1}^{B_{\text{perm}}} \mathbf{1}[S_b \geq S_{\text{obs}}]}{B_{\text{perm}} + 1}.$$

Including equality makes ties conservative. Under exchangeability the observed statistic and the randomized statistics have exchangeable ranks, so this is a valid, possibly conservative, test of the family null. Any searched binary contrasts must be included in the same maximum. Any label-dependent fitting or selection must be repeated inside each permutation. Holding only the observed winning map fixed does not adjust for selection. Ordinary row permutations are not justified by clustered or nonexchangeable sampling.

Separately, Result 4b supplies simultaneous population confidence bounds for a fixed family by replacing β with β/m in each map's radius. With probability at least $1 - \beta$, all map bounds hold, and $|S_{\text{plug}} - \Lambda_{\max}|$ is bounded by their largest radius. Result 4b is a statement about the plug-in, so this bound is for S_{plug} and does not transfer to S_{cf} merely because both summarise the same family. We do not apply it to the surveyed cross-fitted values. This confidence statement requires the i.i.d. sampling assumptions of Result 4b. A permutation null alone is not a confidence interval for $\Lambda_{\max}$. These formulas formalize the declared-family protocol. They do not introduce new maps, change frozen survey results, or bound scores outside the declared family.

Bootstrap comparisons and prospective budgets. Paired comparisons resample the same evaluation units for both scores and recompute their AUC difference. The human-sheet ranges are nominal-weight sensitivity summaries. Re-DocRED resamples evaluation documents and uses the 2.5th and 97.5th percentiles of paired differences. Its overlapping splits are not independent replications. A point difference, the mean of bootstrap differences, and the midpoint of a percentile interval are distinct summaries. The prospective budget curve resamples with replacement from the fitting errors at *every* budget. Budgets at or above the original error count are still bootstrap replicates, not newly observed annotations. The resulting spread is conditional on that finite error pool and does not measure fresh-annotation variability in either direction. A log-error ratio at a fixed budget is not a sample-size multiplier.

F.1 Diagnostic survey and support examples

A support conflict makes us decline a point summary, as for the sparse full-signature map with estimate 0.054 and floor 0.296. Sample splitting alone supplies no general unbiasedness or coverage guarantee (Chernozhukov et al., 2018; Andrews & Guggenberger, 2009). Result 4b gives a potentially vacuous finite-sample bound for the plug-in on a fixed map under i.i.d. unweighted sampling.

Using existing labels and cell keys, we compute Λ across 10 construction-and-cell-map rows without new annotation (Appendix K). Real error processes carry substantial cell structure: $\widehat{\Lambda}_{\text{cf}} = 0.183$ on CoNLL++'s hand-corrected labels, 0.390 on Re-DocRED's human-removed triples and 0.202 on our own annotated claims. Holding graph, anomaly ratio and seeds fixed and changing only the corruption takes WN18RR from 0.226 to a negative estimate reported as 0.000, and FB15k-237 from 0.221 into $[0.012, 0.041]$, where the guard declines to report a point.

Artifact-removal constructions and negative cases. For NLI, 0.179 on MNLI (Williams et al., 2018) falls to 0.000 on its lexical-overlap counterpart HANS. For fact verification, FEVER's (Thorne et al., 2018) 0.070 falls to 0.000 on FEVER-Symmetric. The two differ in evidential value. The

scheme is selected on the original benchmark's Λ alone, and on HANS it yields one cell, so that zero follows from a constant feature rather than from a measurement, leaving FEVER-Symmetric the informative collapse on 708 rows. Across ANLI's increasingly adversarial rounds R1–R3 (Nie et al., 2020) the decline is not monotone, contrary to the prespecified prediction. VitaminC (Schuster et al., 2021), which pairs claims with near-identical supporting and refuting evidence, falls only partly because its construction changes the claim–evidence relation, whereas our diagnostic retains claim features (Appendix D).

G A PROSPECTIVE GENERALIZATION TEST ON SUMMARY GROUNDING ERRORS

Design and isolation. We froze a local protocol before downloading RAGTruth (Niu et al., 2024), then froze the implementation and fitted heads before evaluating its official test responses. This is a prospective local protocol, not external preregistration. We retain Summary responses with upstream quality marked good, across all six generators. A response is an error if it has any annotated hallucination span, including spans annotated as unsupported but potentially true. The alternative definition omits those spans. Neither definition is chosen using test outcomes. Sources with identical normalized text or token-five-gram Jaccard similarity at least 0.80 form transitive components. Training responses connected to a test component are removed. A deterministic source-component hash partitions the remaining training data into fitting, selection and independent confirmation roles in proportions $0.60/0.20/0.20$. There are 478, 161 and 152 development components respectively. Three responses fail the upstream quality filter. The official test contains 900 responses in 150 components, with 204 errors. Source grouping reduces direct leakage but does not establish independence of related news events.

Fixed representation. For a context–response pair $x = (c, r)$, let $\phi(x) \in \mathbb{R}^{15}$ be a fixed feature vector. We split both texts at sentence punctuation followed by whitespace, or at newlines. For each of the first twelve nonempty response sentences, we retrieve up to three context sentences by cosine similarity of binary token sets after removing a fixed function-word list. Ties use source position, and retrieved sentences are concatenated in source order. Frozen RoBERTa-large-MNLI (Liu et al., 2019) evaluates each pair with at most 384 premise tokens, 124 hypothesis tokens and four special tokens. Minimum, mean and maximum contradiction, neutral and entailment probabilities give nine coordinates. The remaining six are $\log(1 + \text{count})$ for response tokens, context tokens and all response sentences, and the fractions of unique response unigrams, bigrams and numeric tokens absent from the context. Empty denominators give zero. Lexical tokens are lower-cased Unicode word sequences, distinct from the model tokenizer's subword tokens used for the 384/124 limits. Numeric tokens contain at least one digit. The primary representation contains neither generator identity nor human labels. The model is not fine-tuned here. Its pretraining may contain the source news, so these are held-out source groups for our fitting procedure, not a claim of unseen text in pretraining. The release pins data and checkpoint revisions, model-file hashes, donor identities and all truncation counts.

Empirical laws and fitted scores. Let n be the number of fitting responses, x_i their true context–response pairs, $y_i \in \{0, 1\}$ their error labels and $n_e = \sum_{i=1}^{n} y_i$. Writing δ_x for a unit point mass at x, the fitted pool and auxiliaries are

$$\widehat{P} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}, \qquad \widehat{Q}_m = \frac{1}{n_e}\sum_{i=1}^{n} y_i \delta_{x_i}, \qquad \widehat{Q}_g = \frac{1}{n}\sum_{i=1}^{n} \delta_{(c_{d(i)}, r_i)}.$$

Here $d(i)$ is a deterministic donor from another source component in the same generator/news-source stratum. The matched law uses 874 observed fitting errors. The generic law keeps each response and replaces its context. The pool includes errors, with labels hidden from the pool side of the fit. For dilution parameter $\lambda \in \{0.05, 0.10, \ldots, 1\}$, define

$$\widehat{Q}_{g,\lambda} = (1 - \lambda)\widehat{P} + \lambda \widehat{Q}_g.$$

Mixture weights are exact, without resampling. Let $z(x)$ denote $\phi(x)$ standardized using only the fitting pool. For each auxiliary Q, the linear score $s_Q(x) = w_Q^\top z(x) + b_Q$ minimizes

$$\frac{1}{2}\mathbb{E}_{\widehat{P}} \log(1 + e^{w^\top z(X) + b}) + \frac{1}{2}\mathbb{E}_{Q} \log(1 + e^{-w^\top z(X) - b}) + \frac{0.01}{2}\|w\|_2^2.$$

Score	AUC	95% interval	P@5%
Generic mismatches	0.6982	[0.6599, 0.7369]	0.5556
Human-error matched	0.7202	[0.6833, 0.7577]	0.5333
Permuted-label matched	0.7171	[0.6803, 0.7545]	0.5556
Supervised, same labels	0.7200	[0.6831, 0.7576]	0.5333
Length + generator/source prior	0.7999	[0.7679, 0.8331]	0.5778
Frozen NLI baseline	0.6552	[0.6168, 0.6944]	0.4000

Table 6: All prespecified RAGTruth test scores. Intervals resample source components with fitted heads held fixed. P@5% uses fractional boundary ties.

Generator	Rows	Errors	Generic AUC	Matched AUC
gpt-3.5-turbo-0613	150	4	0.7774	0.7414
gpt-4-0613	150	6	0.6852	0.6759
llama-2-13b-chat	150	33	0.6384	0.6408
llama-2-70b-chat	150	24	0.6194	0.6587
llama-2-7b-chat	150	51	0.6564	0.6730
mistral-7B-instruct	150	86	0.6648	0.6855

Table 7: Every generator in the held-out RAGTruth Summary split. These subgroup AUCs are descriptive and do not replace the pooled primary contrast.

Thus larger scores indicate auxiliary membership and are used to rank errors. The intercept b is unpenalized. Fits use float64 arithmetic and require gradient infinity norm at most 10^{-6}. Each head is fitted once.

Failed hardness confirmation. Selection minimizes the absolute difference between the two fitted heads' own auxiliary-versus-pool AUCs, requiring both to be at least 0.60 and their difference at most 0.03. For a fixed generic head, dilution AUC equals $1/2+\lambda\{\mathrm{AUC}(Q_g, P)-1/2\}$. The selected dilution is 0.25. Independent confirmation uses 5,000 coupled source-component bootstrap draws. A generic mismatch receives the product of donor and recipient multiplicities, while true pairs receive their own component multiplicity. The paired 90% difference interval $[-0.0280, 0.0133]$ lies within $[-0.03, 0.03]$, but the matched arm's one-sided 95% lower AUC bound is 0.5967, below the required 0.60 (generic: 0.6245). The prespecified conjunction therefore fails. No alternative dilution is tried and no matched-hardness test claim is made. The protocol explicitly retains the primary undiluted comparison after this failure.

Inference and controls. The primary statistic is test AUC of s_{Q_m} minus test AUC of s_{Q_g}. AUC gives half credit to tied scores. Its paired percentile interval uses 5,000 valid source-component bootstrap draws, holding every fitted head fixed. The positive criterion requires the 95% interval to exclude zero, and the strong criterion additionally requires a point difference of at least 0.03. These thresholds preceded test evaluation. P@5% inspects exactly five percent of the response mass, distributing any boundary tie fractionally and uniformly. The same-label supervised control uses all fifteen features and the same fitting annotations. The prior control uses three log-length features plus generator and news-source indicators. Permuted matching shuffles fitting labels within those strata once using a fixed seed. It is a negative control, not a permutation test. The training-free NLI score is one minus minimum entailment probability. Per-generator results appear in Table 7. All control comparisons and per-generator results are descriptive. Intervals exclude fitting uncertainty, annotation error and residual event dependence.

Outcome and scope. The sentence cap affects 17 test responses. Matched P@5% is slightly lower than generic P@5% (Table 6). The primary difference is 0.022 with 95% interval $[0.005, 0.040]$: a positive direction, but below the fixed strong-effect threshold. Matched and same-label supervised AUCs are almost identical. Permuted matching also nearly ties the matched head, and the pooled prior scores higher. Table 8 shows that the macro within-generator advantage is smaller than the pooled advantage. The prior's pooled strength does not imply strong discrimination within each generator. These results do not isolate semantic fidelity from stratum composition or establish a

Score	Macro generator AUC	Alternative-label AUC
Generic mismatches	0.6736	0.6953
Human-error matched	0.6792	0.7140
Permuted-label matched	0.6391	0.7122
Supervised, same labels	0.6806	0.7138
Length + generator/source prior	0.5936	0.7932
Frozen NLI baseline	0.6177	0.6547

Table 8: Prespecified descriptive sensitivity summaries. Macro AUC averages the six within-generator AUCs equally. Alternative-label AUC excludes spans marked unsupported but potentially true from the error definition.

benefit over supervised learning. The alternative label definition and every generator are retained without replacing the primary comparison. The released replay recomputes all statistics from stored scores. It does not refit the heads or regenerate frozen NLI features.

G.1 Exploratory mechanism analysis after the primary evaluation

This subsection was designed after seeing the primary results. It reuses the frozen scores, fits no new model and is not a confirmatory replication.

Information retained by within-stratum permutation. Partition the fitting rows into strata g, one for each generator and news source. Let n_g be a stratum's size, e_g its number of labelled errors, $\widehat{\pi}_g = e_g/n_g$, and $\widehat{\pi} = n_e/n$ the pooled error rate. Under uniform permutation of labels within each stratum, a row in stratum g has probability e_g/n_g of being selected for the matched auxiliary. For row-indexed empirical measures, the expected permuted auxiliary therefore has

$$\mathbb{E}_{\mathrm{perm}}\widehat{Q}_{\mathrm{perm}} = \sum_g \frac{e_g}{n_e}\widehat{P}_g, \qquad \frac{d\,\mathbb{E}_{\mathrm{perm}}\widehat{Q}_{\mathrm{perm}}}{d\widehat{P}}(i) = \frac{\widehat{\pi}_{g(i)}}{\widehat{\pi}},$$

where $\widehat{P}_g$ is uniform over the fitting rows in stratum g, and $g(i)$ denotes row i's stratum. The total error count n_e is fixed by the permutation. Thus the control removes the association with error labels *within* strata but preserves their different prevalences. After mapping rows to features, its density ratio is the conditional expectation of this stratum ratio given the retained features. Generator identity need not be an explicit feature for style or length to carry some of that information. This is an exact sampling-law identity, not a claim that the expectation of fitted scores equals the score fitted to the expected law. The actual control uses one fixed permutation and a restricted regularized linear head.

Separating within- and cross-stratum comparisons. For independent error and correct draws $X^+ \sim E$, $X^- \sim C$, let $G^+ = g(X^+)$ and $G^- = g(X^-)$, and set $\omega = \Pr(G^+ = G^-)$. Define $A_=(s)$ and $A_{\neq}(s)$ as expected half-tie comparison credit conditional on equal and unequal strata, respectively. Then

$$\mathrm{AUC}(s) = \omega A_=(s) + (1-\omega)A_{\neq}(s).$$

This is the law of total expectation. In the unweighted empirical sample, $\omega = \sum_g n_{1g}n_{0g}/(n_1 n_0)$, where n_{yg} counts label-y rows in stratum g and $n_y = \sum_g n_{yg}$. For a comparison of two fixed scores, the same weights decompose their AUC difference exactly. These are pair-mass-weighted conditional AUCs, distinct from the equal-generator macro AUC in Table 8. A zero-mass conditional term is undefined and its contribution is zero. Table 9 gives the exhaustive decomposition for both generator and generator/source partitions, across every control. Most of the matched-versus-generic pooled gain arises across generators: 0.0199 compared with a within-generator contribution of 0.0021. Conversely, the prior's pooled advantage comes from cross-generator comparisons, while matched training has the higher within-generator AUC. Neither observation alone establishes a causal semantic mechanism or broad transfer.

Agreement of matched and supervised scores. At population level, matched contrastive training with $Q = E$ orders by $dE/dP = \eta/\pi$, whereas supervised error classification orders by $\eta/(1-\eta)$,

Stratum	Matched minus	Within	Across	Total
Generator	Generic	+0.0021	+0.0199	+0.0220
Generator	Permuted	+0.0044	-0.0013	+0.0031
Generator	Supervised	+0.0001	+0.0001	+0.0002
Generator	Prior	+0.0159	-0.0956	-0.0797
Generator	NLI	+0.0063	+0.0587	+0.0650
Gen.+source	Generic	+0.0015	+0.0206	+0.0220
Gen.+source	Permuted	+0.0025	+0.0005	+0.0031
Gen.+source	Supervised	+0.0001	+0.0001	+0.0002
Gen.+source	Prior	+0.0088	-0.0885	-0.0797
Gen.+source	NLI	+0.0027	+0.0623	+0.0650

Table 9: Exploratory decomposition of matched-minus-control AUC differences. Within and cross columns already include their pair-mass weights and sum to the total, up to rounding. No subgroup selection or refitting is performed.

Stratum	Matched minus	AUC difference	95% interval
Generator	Generic	+0.0157	[-0.0089, +0.0412]
Generator	Permuted	+0.0331	[+0.0037, +0.0623]
Generator	Supervised	+0.0007	[-0.0042, +0.0053]
Generator	Prior	+0.1197	[+0.0485, +0.1859]
Generator	Nli	+0.0477	[+0.0020, +0.0951]
Gen.+source	Generic	+0.0192	[-0.0077, +0.0460]
Gen.+source	Permuted	+0.0334	[+0.0017, +0.0662]
Gen.+source	Supervised	+0.0012	[-0.0044, +0.0066]
Gen.+source	Prior	+0.1152	[+0.0426, +0.1859]
Gen.+source	Nli	+0.0355	[-0.0092, +0.0794]

Table 10: Post-outcome conditional AUC differences and descriptive source-cluster intervals. These conditional differences are not multiplied by the within-stratum pair fraction, unlike the contributions in Table 9.

where $\eta = \Pr(Y = 1 \mid X)$. Both are increasing functions of η. Thus agreement of ideal rankings is expected from the density-ratio identity, not evidence of a new advantage over supervision. It does not require finite-sample regularized linear fits to agree.

Exploratory uncertainty, not a replacement primary endpoint. Conditioning on pairs from the same generator, matched minus generic AUC is 0.0157, with descriptive 95% interval $[-0.0089, 0.0412]$. Matched minus permuted AUC is 0.0331 with interval $[0.0037, 0.0623]$. Thus pooled near-equality does not establish absence of within-generator label information. These post-outcome intervals use 2,000 source-component bootstrap draws, with all heads fixed and the conditional pair-mass weights recomputed in each draw. They are unadjusted and exploratory. Table 10 retains every control and both partitions. The generic comparison remains unresolved, and these analyses neither replace the original primary endpoint nor establish a semantic causal effect.

H Prospective QA-domain study with matched stratum marginals

The QA protocol was locally frozen before QA development outcomes were inspected. It fixes a conditional target and equal auxiliary stratum marginals, rather than selecting a favourable subgroup after evaluation. This is task/source generalization within RAGTruth, with the same annotation pipeline and generator families. It is not a new annotation campaign.

Sources, exclusion and isolation. We retain all good-quality QA responses across the six generators. The primary error label is any annotated hallucination span, including statements that are unsupported but potentially true. The alternative omits that span category. Evidence consists of the retrieved passages, excluding the question itself. The data revision and frozen NLI checkpoint are

the same as in Appendix G, and all fifteen features, retrieval rules, sentence/token caps and numerical fitting settings are unchanged. QA sources are connected by normalized passage identity or fivegram Jaccard at least 0.80. Any component connected to a Summary source is excluded. Official QA test components are reserved, and any linked training responses are removed. A source-component hash assigns remaining training components to fitting, development and reserved evaluation in fractions 0.50, 0.20 and 0.30. All sibling responses stay together. Reserved and official test sources are then combined for the frozen evaluation. No target label count or feature is used for fitting or the power decision. Upstream quality filtering removes 143 incorrect refusals and 24 truncated responses.

There are 411 fitting components with 2388 responses and 768 errors, and 184 development components. Every fitting stratum contains both classes. The target has 394 components, 2305 responses and 612 errors. Source linking screens direct reuse but does not prove independence of related passages or absence of pretraining exposure. The NLI checkpoint is not fine-tuned here.

Balanced laws and fitted heads. For fitting stratum g (generator plus source corpus), let n_g and e_g denote its row and error counts, and $n = \sum_g n_g$. The pool is uniform over fitting rows. Each error in the balanced matched auxiliary receives mass $(n_g/n)/e_g$, making its stratum marginal exactly n_g/n. Generic mismatches keep each response but pair it with a different component's passages in the same stratum, preserving the same marginal. Balanced permutation shuffles the labels within strata once with a fixed seed and uses the same weights. Its expected auxiliary law equals the pool, as proved after Result 5e. The three contrastive heads use the balanced logistic objective from Appendix G, ridge 0.01, unpenalized intercept and pool-only standardization. No head is selected using an audit outcome.

Controls use the same fitting annotations: supervised logistic and a length/stratum prior assign each class the same stratum masses n_g/n. Unbalanced matching, the raw NLI score and two unchanged Summary-trained heads are also reported. The Summary heads retain their original standardizer and receive no QA fitting labels. Their evaluation is descriptive cross-task transfer. The global score cannot be assumed to attain the conditional oracle merely because group masses agree: the feature-only linear class may be misspecified (Result 5e). Cross-task claims also require accounting for pool shift (Result 5f).

Power gate and pre-target freeze. The fixed feasibility floors require at least 100 fitting, 60 development and 250 evaluation components, at least 40 errors and correct responses in fitting, at least 20 of each on development, and both classes in every fitting stratum. All floors pass. Before target access, 2,000 development source-component bootstrap draws estimate the standard deviation $\widehat{\sigma}_j$ of each of the two conditional AUC differences. For development and target component counts G_d, G_t, the planning calculation of the target standard error SE and of the minimum effect detectable with 80% power, MDE_{80}, is

$$\widehat{\mathrm{SE}}_{t,j} = \widehat{\sigma}_j \sqrt{G_d/G_t}, \qquad \widehat{\mathrm{MDE}}_{80,j} = (z_{0.9875} + z_{0.8})\widehat{\mathrm{SE}}_{t,j},$$

where z_u is the standard-normal u quantile. The gate requires both MDEs at most 0.05. It uses variability, not the observed effect's sign or magnitude, and is a planning approximation, not a power guarantee at the 0.03 strong-effect threshold. The generic and permutation MDEs are 0.0380 and 0.0480, so the gate passes. Development matched-minus-generic AUC is negative (-0.0117), which does not change any head or decision rule. The implementation, heads, input/model hashes and power decision were frozen in a local commit before target feature extraction or label analysis.

Primary metric and inferential rule. For independent error/correct pairs, condition on both endpoints sharing a stratum. Empirical conditional AUC divides the total within-stratum half-tie comparison credit by $\sum_g n_{1g} n_{0g}$, where n_{yg} counts class y in stratum g. It uses the pair-mass weights of Result 5e, not uniform generator weights or pooled AUC. The two co-primary differences are balanced matched minus generic and balanced matched minus balanced permutation. Each has a 97.5% percentile interval from 5,000 coupled source-component bootstrap draws, giving Bonferroni control for the two interval-based directional claims under the bootstrap approximation. Conditional pair weights are recomputed in every draw. Positive confirmation requires both lower bounds above zero. Strong confirmation additionally requires both point effects at least 0.03. All heads remain fixed. No hardness equivalence is asserted. Neither control differences nor transfer results can replace a failed co-primary comparison.

Table 11 gives all scores, Table 12 gives the generator strata, and Table 13 reports label and aggregation sensitivities.

Score	Conditional AUC	95% interval	Pooled AUC	P@5%
Balanced matched	0.7317	[0.7031, 0.7603]	0.7913	0.7223
Generic mismatches	0.6839	[0.6530, 0.7144]	0.7446	0.5488
Balanced permutation	0.6199	[0.5901, 0.6506]	0.6865	0.3991
Balanced supervised	0.7444	[0.7159, 0.7730]	0.8044	0.7831
Balanced length/stratum prior	0.6566	[0.6230, 0.6902]	0.6495	0.5813
Unbalanced matched	0.7614	[0.7342, 0.7892]	0.8223	0.8503
Frozen NLI	0.6274	[0.5982, 0.6560]	0.6895	0.3839
Summary-trained matched	0.7007	[0.6706, 0.7309]	0.7693	0.5054
Summary-trained generic	0.6887	[0.6592, 0.7184]	0.7520	0.3059

Table 11: Every predeclared QA score. Conditional AUC is the primary metric. Pooled AUC and P@5% are descriptive. Intervals condition on the fitted heads.

Generator	Rows	Errors	Matched	Generic	Permuted
gpt-3.5-turbo-0613	374	24	0.8261	0.7513	0.7404
gpt-4-0613	383	18	0.5855	0.5559	0.5148
llama-2-13b-chat	393	135	0.7146	0.6963	0.6030
llama-2-70b-chat	392	121	0.7542	0.6755	0.6893
llama-2-7b-chat	391	184	0.6981	0.6950	0.5143
mistral-7B-instruct	372	130	0.7734	0.6744	0.6841

Table 12: All QA generator strata. These subgroup AUCs do not replace the prospective conditional comparison.

Score	Macro stratum AUC	Alternative conditional AUC
Balanced matched	0.7253	0.7229
Generic mismatches	0.6747	0.6949
Balanced permutation	0.6243	0.6041
Balanced supervised	0.7362	0.7380
Balanced length/stratum prior	0.6476	0.6322
Unbalanced matched	0.7505	0.7406
Frozen NLI	0.6321	0.6100
Summary-trained matched	0.6903	0.7009
Summary-trained generic	0.6802	0.6981

Table 13: All QA sensitivity summaries. Alternative labels exclude unsupported but potentially true spans. Macro AUC weights informative strata equally.

Outcome and limitations. Both co-primary comparisons pass: matched minus generic is 0.048 with adjusted interval $[0.021, 0.077]$, and matched minus balanced permutation is 0.112 with interval $[0.073, 0.151]$. Both point effects exceed 0.03, satisfying the fixed strong criterion. This is an advantage over those two auxiliaries at fixed stratum composition, not a claim of equal hardness or superiority to all label-matched learning. Supervised conditional AUC is 0.744 and unbalanced matching reaches 0.761, both higher than balanced matching's 0.732. The unchanged Summary heads give 0.7007 (matched) and 0.6887 (generic), a descriptive transfer comparison without a new superiority test.

The twelve-sentence cap affects 572 target responses. Labels cover the entire answer, so later unsupported content can be missed by the NLI coordinates, although the length/lexical coordinates use the full response. The cap remained fixed throughout evaluation. The combined reserved/official target differs from the official-test-only benchmark and should not be compared to published benchmark results. Related-passage dependence, annotation error, checkpoint pretraining and fitting uncertainty

remain outside the intervals. The result does not establish generalization to new generator families or an independent annotation campaign. All sensitivity tables retain the original primary label definition and decision rule.

H.1 POST-HOC FITTING-SAMPLE ROBUSTNESS

This analysis of fitting uncertainty was specified after the primary outcomes were known. The original QA protocol, heads, scores, co-primary intervals and strong-decision rule remain unchanged. This is a post-hoc robustness analysis, not a new prospective test or replication. Before examining the new bootstrap outcomes, we fixed the resampling rule, 5000 replicates, five fitted arms and all reported interval levels.

Let G_f and G_t denote the numbers of fitting and evaluation source components. In replicate b, independently draw component multiplicities

$$(M^f_{b1}, \dots, M^f_{bG_f}) \sim \mathrm{Multinomial}\big(G_f, (1/G_f, \dots, 1/G_f)\big),$$
$$(M^t_{b1}, \dots, M^t_{bG_t}) \sim \mathrm{Multinomial}\big(G_t, (1/G_t, \dots, 1/G_t)\big).$$

Every sibling response carries its component's multiplicity. On the replicated fitting rows, recompute the pool mean and standard deviation, stratum marginals and balanced auxiliary weights, then refit the matched, generic, permutation, supervised and unbalanced heads. Features, regularizer, numerical tolerances, donor assignments and the original permuted labels stay fixed. All five arms share the same two component draws. No evaluation feature or label enters fitting. The unresampled implementation reproduces all five original coefficients and evaluation scores to absolute tolerance 10^{-10}.

For a refitted score s_b and evaluation row i, write $w_{bi} = M^t_{b,g(i)}$, where $g(i)$ is its source component, h_i its generator/source stratum, x_i its frozen feature vector, and $y_i \in \{0, 1\}$ its error indicator. The joint replicate statistic is

$$\widehat{A}_b(s_b) = \frac{\sum_{i:y_i=1} \sum_{j:y_j=0} w_{bi} w_{bj} \mathbf{1}[h_i = h_j] \big(\mathbf{1}[s_b(x_i) > s_b(x_j)] + \frac{1}{2}\mathbf{1}[s_b(x_i) = s_b(x_j)]\big)}{\sum_{i:y_i=1} \sum_{j:y_j=0} w_{bi} w_{bj} \mathbf{1}[h_i = h_j]}.$$

Setting all evaluation weights to one gives the fitting-only sensitivity statistic. The two contrasts use coupled differences within each replicate. We report their 1.25th and 98.75th percentiles, mirroring the original two-comparison interval levels, and 95% percentile intervals for each arm. Original point estimates remain the reported effects. There is one independent evaluation draw per fitting draw, rather than repeated inner evaluation draws. All replicates completed without missing-class or convergence failures.

Score or contrast	Original	Level	Fitting only	Fitting + evaluation
Balanced matched	0.7317	95%	[0.6866, 0.7403]	[0.6730, 0.7531]
Generic mismatches	0.6839	95%	[0.6780, 0.6891]	[0.6524, 0.7143]
Balanced permutation	0.6199	95%	[0.5249, 0.6583]	[0.5175, 0.6646]
Balanced supervised	0.7444	95%	[0.7148, 0.7521]	[0.6995, 0.7670]
Unbalanced matched	0.7614	95%	[0.7493, 0.7636]	[0.7286, 0.7852]
Matched minus generic	0.0478	97.5%	[-0.0023, 0.0606]	[-0.0155, 0.0752]
Matched minus permutation	0.1118	97.5%	[0.0419, 0.2030]	[0.0316, 0.2172]

Table 14: Post-hoc QA source-bootstrap robustness. Fitting-only intervals hold the evaluation sample fixed. Joint intervals resample both source samples. Interval levels are explicit for each row. All five fitted arms are retained.

Outcome and interpretation. Table 14 reports the fitting-only and joint resampling results. The joint matched-minus-generic interval is $[-0.015, 0.075]$ and includes zero. The matched-minus-permutation interval is $[0.032, 0.217]$ and remains positive. Thus the original fixed-head strong result does not establish fitting-stable superiority over generic mismatches. The post-hoc intervals qualify the interpretation of the original result without replacing its primary endpoint.

The bootstrap propagates fitting-source sampling through standardization, weighting and head estimation. It conditions on the feature extractor, donor context bank, mismatch assignments and one permutation. In particular, generic pairs may share donor contexts, and this analysis does not independently regenerate corruptions. It also omits protocol-selection, annotation, representation and new-generator uncertainty. Percentile coverage is approximate. All refits and statistics replay from anonymous numeric inputs. Independent pair-credit calculations check every replicate, and a separate Newton solver checks the original fit and twenty predetermined resamples.

I FROZEN QA-HEAD TRANSFER TO STRUCTURED-RECORD GENERATION

Transfer question and protocol status. After the QA result, we froze a separate transfer protocol before analyzing Data2txt outcomes. The question is whether the existing balanced-matched QA head outperforms its generic-mismatch counterpart on a different task and source corpus without target fitting. This sequential follow-up is locally frozen, not externally pre-registered or chosen independently of the QA result. Its single primary contrast has its own interval. We claim no familywise significance across all sequential studies.

Data, serialization and isolation. We retain all good-quality Data2txt responses in the same pinned RAGTruth revision, including both upstream splits because neither enters fitting. The six generators produce business descriptions from structured Yelp records. Every scalar source field is serialized in lexicographically sorted dictionary-key order, with dot-separated paths, zero-based list indices, lower-case Boolean values and *unknown* for null. String values are preserved. No field is selected using outcomes. Normalized context identity and fivegram Jaccard at least 0.80 define transitive source components. Any component connected to a prior Summary or QA source is excluded. No such exclusions occur. Three upstream truncated responses are removed. The retained set has 6195 responses from 1033 components. The primary error label includes any annotated span, including unsupported but potentially true statements and misinterpretations of null fields. Alternative labels omit only unsupported but potentially true statements. There are 4251 primary errors.

Metadata gate before target access. The protocol requires at least 500 components, nonempty responses, valid source records and all six generators. Planning uses the existing QA-development standard deviation $\widehat{\sigma}_{\mathrm{QA}}$ of matched-minus-generic conditional AUC, development count $G_d = 184$ and the new metadata-only component count $G_t = 1033$:

$$\widehat{\mathrm{SE}}_{\mathrm{transfer}} = 2\widehat{\sigma}_{\mathrm{QA}}\sqrt{G_d/G_t}, \qquad \widehat{\mathrm{MDE}}_{80} = (z_{0.975} + z_{0.8})\widehat{\mathrm{SE}}_{\mathrm{transfer}}.$$

Here z_u is the standard-normal quantile. The fixed factor two is an assumed inflation for task shift, not a proven variance bound. The MDE is 0.0426, below the fixed 0.05 ceiling. This is not a guarantee of power at a 0.03 effect. Protocol, code, metadata/group hashes and QA-head hashes were committed before new feature extraction or label summaries. No gate or scoring rule was adapted after inspecting target outcomes.

Unchanged models and inference. All nine QA-stage scores are applied with their existing coefficients and standardizers, including the two older Summary-trained transfer scores. There is no target fitting, calibration, label balancing or hyperparameter selection. The fifteen-feature representation, frozen NLI checkpoint, top-three premise retrieval, twelve-response-sentence cap and token limits are unchanged. Core heads receive no generator or source indicator. The prior control's unseen Yelp indicator is zero under the existing encoding fitted on MS MARCO (Bajaj et al., 2016) sources. This is a limitation of that descriptive control.

The primary metric is the conditional AUC from Appendix H, using error/correct pairs from the same generator/source stratum. The only primary contrast is balanced matched minus generic. We use 5,000 coupled source-component bootstrap draws, recompute pair-mass weights and give half credit to ties. The fixed random seed is 2020421875. A two-sided 95% percentile interval determines positive confirmation by a lower bound above zero. Strong confirmation additionally requires a point difference of at least 0.03. All other scores and sensitivity metrics are descriptive and cannot replace this contrast. Fitting uncertainty is outside these fixed-head intervals.

Table 15 reports all scores. Table 16 separates generator strata, and Table 17 checks alternative labels and aggregation.

Score	Conditional AUC	95% interval	Pooled AUC	P@5%
Balanced matched	0.5545	[0.5333, 0.5752]	0.5815	0.8999
Generic mismatches	0.5767	[0.5558, 0.5971]	0.7339	0.9064
Balanced permutation	0.5535	[0.5331, 0.5736]	0.6053	0.7256
Balanced supervised	0.5702	[0.5488, 0.5908]	0.6037	0.8870
Balanced length/stratum prior	0.5840	[0.5633, 0.6043]	0.3126	0.2647
Unbalanced matched	0.6104	[0.5899, 0.6308]	0.6621	0.9322
Frozen NLI	0.5412	[0.5212, 0.5620]	0.5414	0.7232
Summary-trained matched	0.6037	[0.5832, 0.6242]	0.7408	0.9387
Summary-trained generic	0.5814	[0.5604, 0.6019]	0.7333	0.9128

Table 15: All unchanged QA-stage scores on Data2txt. Only the conditional matched-minus-generic contrast is primary. Other scores and metrics are descriptive.

Generator	Rows	Errors	Matched	Generic	Permuted
gpt-3.5-turbo-0613	1033	272	0.5680	0.6037	0.5601
gpt-4-0613	1033	290	0.4688	0.5445	0.5829
llama-2-13b-chat	1032	982	0.5310	0.5528	0.5353
llama-2-70b-chat	1033	863	0.6172	0.6414	0.5748
llama-2-7b-chat	1032	887	0.5793	0.5526	0.5096
mistral-7B-instruct	1032	957	0.6158	0.5235	0.4941

Table 16: Every Data2txt generator stratum under the unchanged QA heads.

Score	Macro stratum AUC	Alternative conditional AUC
Balanced matched	0.5634	0.5503
Generic mismatches	0.5697	0.5663
Balanced permutation	0.5428	0.5503
Balanced supervised	0.5828	0.5687
Balanced length/stratum prior	0.6052	0.5905
Unbalanced matched	0.6213	0.6108
Frozen NLI	0.5534	0.5357
Summary-trained matched	0.6058	0.5985
Summary-trained generic	0.5728	0.5714

Table 17: Every Data2txt sensitivity summary, with no replacement of the primary label.

Outcome. The primary comparison is negative: balanced matched conditional AUC is 0.554 against generic 0.577, giving -0.022 with interval $[-0.041, -0.004]$. Neither positive nor strong confirmation passes. The unchanged supervised head reaches 0.5702 and unbalanced matching reaches 0.6104, both above balanced matching. Alternative labels also retain its negative difference from generic. Higher descriptive control scores do not establish the prespecified transfer hypothesis. The inference is an observed failure to transfer this advantage, not proof that balancing itself caused the failure or that the same ordering must hold on every new task.

Scope. This is a new task, source corpus and collection of source records, but the annotation pipeline and generator families remain shared. The feature cap affects 912 responses. Labels cover the complete response. The fixed source serialization and premise retrieval can lose semantic information. Source pretraining overlap, annotation error and dependence beyond the screened components remain possible. Result 5f explains why error-law fidelity alone does not guarantee transfer across changing pools. Its unknown population discrepancies are not estimated here. The frozen-score supplement independently recomputes all metrics, the primary interval and planning gate with source-pair matrices. This verifies calculations, not reannotation or full model inference.

J CROSSED INTERACTION ON TWO HUMAN ERROR PROCESSES (FRANK)

Question and independence. Every other crossed comparison evaluates on constructed errors or on one human error process. FRANK (Pagnoni et al., 2021) supplies two human error processes in one benchmark, annotated by its own annotators on summaries from nine pre-LLM systems over the news summarization datasets CNN/DM (Hermann et al., 2015; Nallapati et al., 2016) and XSum (Narayan et al., 2018), with no link to RAGTruth's annotation pipeline or generators. The protocol records a directional prediction specified before feature extraction and fitting. It was committed together with the features and heads before error-outcome evaluation. Before specification, only label counts, field semantics and casing statistics were inspected. A label-free spot-check of corruptions tightened the span and casing rules before feature extraction.

Error processes. Correct summaries have factuality one. The entity row contains summaries with an entity error and no out-of-article error (323), the out-of-article row the reverse (414), each against the 810 correct summaries. Other summaries stay in the pool and enter no evaluation row.

Label-free auxiliaries. Both auxiliaries replace one name or number span in a pool summary and differ only in where the replacement comes from. The within-article swap draws a different span of the same class from the same article, mimicking FRANK's entity error. The cross-article swap draws a span from another article of the same dataset and fold that does not occur in the source article, mimicking an out-of-article error. Spans are capitalized runs with a non-sentence-initial occurrence, excluding runs made only of common lower-case WordNet (Miller, 1995) words, and digit tokens. Replacements are lower-cased when the replaced occurrence is lower-case and otherwise retain their source spelling. This prevents inserting a capitalized span into a lower-case occurrence, without establishing absence of all casing cues. Both auxiliaries are built from the same 1948 base summaries and share their stratum marginals exactly. Neither uses an error label.

Heads, folds and estimand. Articles are hashed into two folds. Per fold, the pool is every summary of that fold and two heads separate it from each auxiliary, on the frozen fifteen-feature representation of Appendix G with the article as context, pool-only standardization and the same balanced logistic objective. Heads of one fold score only the other fold. The primary interaction is

$$I = [A_{\text{ent}}(\text{within}) - A_{\text{ent}}(\text{cross})] - [A_{\text{out}}(\text{within}) - A_{\text{out}}(\text{cross})],$$

where $A_r(a)$ is the conditional AUC of auxiliary a on error row r, over error/correct pairs sharing fold and generator stratum, weighted by pair mass as in Result 5e. The prediction is $I > 0$.

Inference with refitted heads. The 2000 article-level bootstrap replicates refit the standardizers and all four heads on the replicated fitting rows and recompute the evaluation statistic, so fitting uncertainty is propagated in the primary interval rather than in a later analysis. Features, spans, donors and corruptions are conditioned on. A positive interaction requires the 2.5th percentile of I above zero, and a full crossover additionally requires both row differences to have 95% intervals above zero.

Outcome. The interaction is 0.206 with 95% interval $[0.121, 0.297]$, so the positive-interaction criterion passes. The full crossover fails. On out-of-article errors the cross-article auxiliary leads by 0.196 $[0.105, 0.288]$. On entity errors the within-article auxiliary leads by only 0.011 $[-0.026, 0.049]$, an interval containing zero (Table 18). The effect is therefore asymmetric, carried by the out-of-article row. On its label-free difficulty proxy, each head's pool-versus-auxiliary AUC on the other fold, the within-article auxiliary is the harder one (0.652 against 0.704). The harder auxiliary is barely better on the errors it was designed to match and much worse on the other process.

Evaluated human errors	Training auxiliary	Conditional AUC	95% interval
Entity errors	Within-article swap	0.6353	[0.5807, 0.6882]
Entity errors	Cross-article swap	0.6244	[0.5708, 0.6744]
Out-of-article errors	Within-article swap	0.5602	[0.4771, 0.6276]
Out-of-article errors	Cross-article swap	0.7558	[0.6789, 0.8132]
Entity row: within minus cross		0.0109	[-0.0264, 0.0492]
Out-of-article row: cross minus within		0.1956	[0.1047, 0.2877]
Interaction (primary)		0.2064	[0.1210, 0.2967]

Table 18: FRANK crossed interaction. Conditional AUC over error/correct pairs sharing fold and generator stratum, with 95% intervals from 2000 article-bootstrap replicates that refit every head. Rows are the human errors evaluated, as in Table 25.

Descriptive checks and limits. All descriptive summaries were fixed in the protocol and are reported in the artifact. Pooled within-fold AUC, precision among the top 5%, and alternative labels counting any entity or out-of-article error agree on the out-of-article row. The entity row is mixed: pooled within-fold AUC favours the within-article head, but with any entity error counted the two tie, and the cross-article head has the higher top-5% precision. The auxiliaries are name and number swaps, so they probe one mechanism of each error type. The systems are pre-LLM summarizers, the frozen NLI model may have seen the news sources in pretraining, the intervals condition on the fixed corruptions and features, and the result does not establish equal hardness or a semantic mechanism.

K Complete benchmark diagnostic results

From labelled cell counts, Λ takes $O(k \log k)$ time by the prefix-sum form (14), and Table 19 works the computation once, on our own annotated claims.

Relation to partial-input baselines. A partial-input baseline trains one model on restricted input and reports that model's accuracy. It measures the performance of that fitted model. The population Λ is the greatest advantage over chance *any* score measurable on the stated partition can attain. Its plug-in counterpart is computed from labelled counts without model training, and Result 3a bounds that empirical quantity from below using cells occupied by only one class. The two quantities describe fitted performance and the maximum attainable on the stated partition, respectively. The surveyed benchmarks do not report the latter. We report cross-fitted sensitivity summaries for these benchmark maps rather than treating the empirical values as population maxima.

Comparison with published partial-input results. The surveyed benchmarks are of two kinds: textual-entailment and natural-language-inference sets (SciTail, GLUE-RTE, SNLI, MNLI, HANS, ANLI), and fact-verification sets (FEVER and FEVER-Symmetric, VitaminC, AVeriTeC, Climate-FEVER). In the partial-input survey of Table 23, over 4 benchmarks, each evaluated with its task's published feature family against a permuted-label null, a shortcut appears in 3. On SciTail (Khot et al., 2018) it is not a discovery, since Poliak et al. (2018) already report a hypothesis-only baseline above majority. What cross-fitted $\widehat{\Lambda}_{\mathrm{cf}} = 0.082\ [0.061, 0.102]$ adds, against a null of 0.020, is a quantity of the construction rather than of one model. Its plug-in counterpart is exact on the rows measured: no score using only those features exceeds AUC 0.586. GLUE-RTE (Wang et al., 2019a) falls within its permutation null, providing no detected shortcut for the tested map. Outside knowledge graphs entirely, SNLI (Bowman et al., 2015) gives cross-fitted $\widehat{\Lambda}_{\mathrm{cf}} = 0.1231$ on a 57-cell map. Its plug-in is exact on the test rows: no cell-only score there exceeds AUC 0.627. A richer hypothesis-only model uses information that map discards and is not bounded by it (Table 23).

CCA benchmark provenance. CCA's repository generates an anomaly mixture during its workflow rather than releasing a fixed benchmark. Its documented generation step does not fix the random seed and depends on BERT (Devlin et al., 2019) and TransE (Bordes et al., 2013) stages. Two executions of the unmodified pipeline on different hosts produced different mixtures. The reported row is one generated instance, identified by checksum in the supplement's provenance package.

cell a	error mass e_a	correct mass c_a
associated with	0.332	0.685
causes	0.348	0.108
part of	0.122	0.126
role of	0.198	0.082
Λ plug-in 0.2029, cross-fitted 0.2018		

Table 19: Λ computed once, on this paper's own corpus. Marginals, plug-in and cross-fitted estimate are all computed from the same 594 resolved rows of the pooled 600, the rest being claims no annotator could settle. Errors concentrate in *causes* and correct claims in *associated with*, and that single opposition supplies almost exactly half of $\sum_{a<b} |D_{ab}| = 2\Lambda$. Plug-in and cross-fitted estimates differ for the reason §5 gives. On a single sample that gap mixes the bias with split-to-split variation and is not a measurement of the bias by itself.

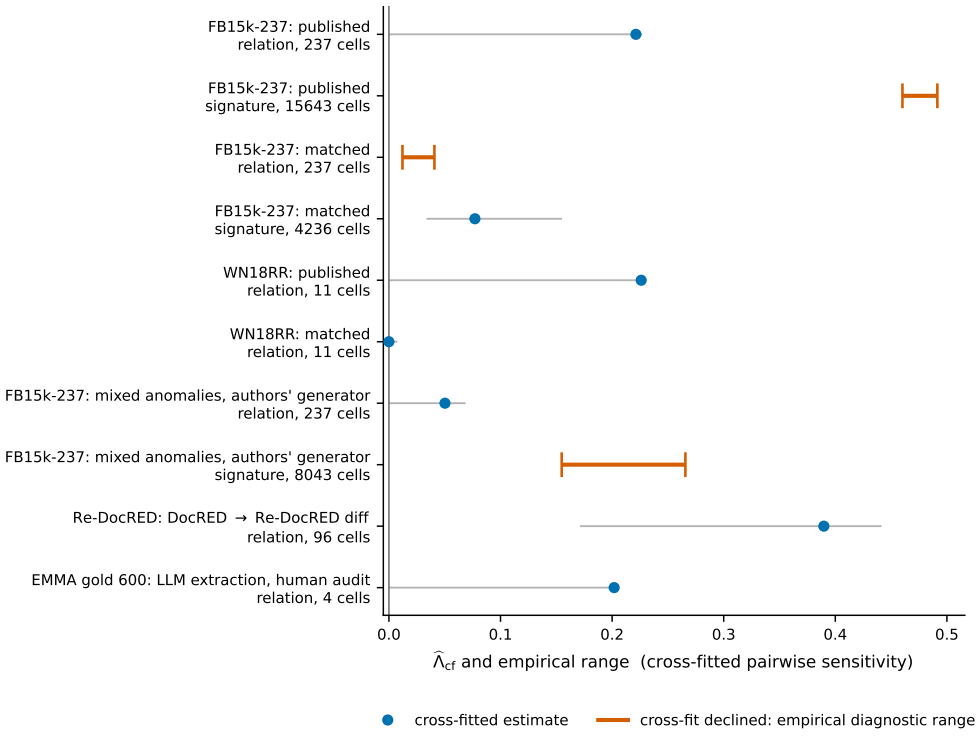

Figure 2: Cross-fitted pairwise sensitivity $\widehat{\Lambda}_{cf}$ across published benchmark constructions and stated cell maps. Points are out-of-fold estimates, not ceilings. What bounds every cell-only score is the population identity of Result 4. Bars show the *empirical diagnostic range* that the estimator reports for every row, not the analytic delta-method interval. Where the support diagnostic declines a point estimate, only this range is shown. Adding $\frac{1}{2}$ gives a mean directional held-out AUC when fitting coverage holds (Appendix F). That condition is sufficient but not necessary, so an uncovered direction does not guarantee the interpretation, and truncating a value for display invalidates it. Table 20 gives the exact values and the seed averaging.

Table 20: Cross-fitted $\widehat{\Lambda}_{\mathrm{cf}}$ across published benchmarks. Constructed benchmarks are re-injected at every seed (the seed draws the benchmark) and the seed average runs over *every* seed, including rows the floor declines: excluding them would bias upward exactly the near-zero constructions the contrast depends on.

benchmark	construction	cells	$\widehat{\Lambda}_{\mathrm{cf}}$
FB15k-237	published (uniform replacement)	237	**0.2213**
FB15k-237	published (uniform replacement)	15,643	*declined*
FB15k-237	matched (within-type replacement)	237	*declined*
FB15k-237	matched (within-type replacement)	4,236	**0.0771**
WN18RR	published (uniform replacement)	11	**0.2261**
WN18RR	matched (within-type replacement)	11	**0.0000**
FB15k-237	mixed anomalies, authors' generator	237	**0.0502** [0.043, 0.057]
FB15k-237	mixed anomalies, authors' generator	8,043	*declined*
Re-DocRED	DocRED → Re-DocRED diff	96	**0.3897** [0.368, 0.411]
EMMA gold 600	LLM extraction, human audit	4	**0.2018** [0.129, 0.274]

at the relation key, changing only the corruption:
WN18RR 0.2261 → 0.0000, and FB15k-237 0.2213 → declined, [0.012, 0.041]

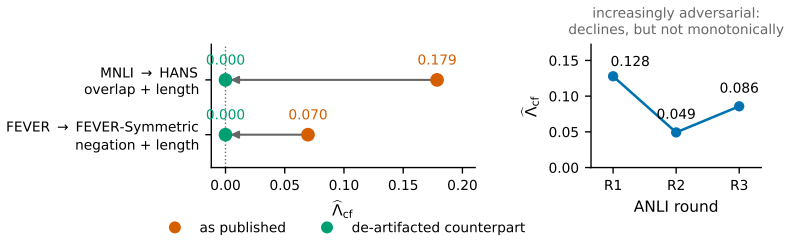

Figure 3: Paired artifact-removal comparison with a prespecified directional prediction. *Left*: each published benchmark beside the counterpart built to remove a known artifact. The cross-fitted estimate collapses to zero on both. Features are the originating papers' own, and pairs are evaluated at *matched* resolution. FEVER reaches 0.151 on its own best partition. *Right*: the ANLI rounds.

Figure 3 compares construction pairs. Table 21 gives their values and distinguishes weak features from unavailable features.

Table 23 requires the following qualifications. **SciTail**'s primary row is its test split, over 2126 rows. The train arm uses no vocabulary fitted to that split and provides a denser descriptive sample. It is not an independent held-out validation. **AVeriTeC** (Schlichtkrull et al., 2023) reduces FEVER's claim-only shortcut to 0.0416 $[0.0069, 0.0762]$ over 2591 claims without removing it, under the reported feature maps and contrasts. This comparison does not isolate claim provenance as a causal explanation. Its own paper reports a no-evidence pipeline (macro F1 0.17 against 0.49 with gold evidence, their Table 4), so partial input is not unexamined there, and what Λ adds is a property of the construction on the specified cell map, estimated out of fold, rather than one model's score. Its 427-row dev arm clears its permutation null but its sensitivity range includes zero, so only the train arm clears both gates, limiting the evidence for this benchmark. **GLUE-RTE** declines at both splits, 0.012 $[-0.021, 0.046]$ over 2490 hypotheses against a null of 0.013, with no separation from the permutation null. The narrow margin supports only a finding of no detected shortcut on the tested maps. Positive Λ values in the other benchmarks summarise cell-label sensitivity on the stated partitions, the plug-in giving the exact attainable maximum on the supplied rows, and do not by themselves make a benchmark unusable. Published partial-input scores, where available, describe individual models.

A regime map measured at a Λ known in closed form puts the collapse below approximately 5 minority samples per cell, where the floor is triggered on 73–100% of draws. At 5 per cell cross-fitting clears the floor while still running 31–38% low, and it is within 3–12% by 20, so the floor

Table 21: The paired de-artifacting test. Features are the originating papers' own, and no feature setting is tuned separately for each benchmark. Pairs are evaluated at *matched* resolution (each arm scored on a scheme both support, chosen by the original's Λ alone). FEVER's two rows differ for that reason alone: the scheme shared with VitaminC is FEVER's own best partition, 0.151, whereas the FEVER-Symmetric pair shares only the coarser joint negation-and-length partition, on which FEVER scores 0.070. WinoGrande (Sakaguchi et al., 2020) and adversarial QA were considered and dropped, the first publishing no artifact feature set and the second having no bipartite-plus-cells structure. HANS's zero follows from a constant map: every HANS pair has full lexical overlap by construction, so the feature is constant and its maximum AUC is one half. The module separates that from a feature that is merely *absent*, where almost everything falls into the residual vocabulary category because the vocabulary does not apply, and only the second is ineligible.

published	de-artifacted counterpart	$\widehat{\Lambda}_{\mathrm{cf}}$	$\widehat{\Lambda}'_{\mathrm{cf}}$	expectation
MNLI	HANS	0.179	0.000	counterpart near zero: met
FEVER	FEVER-Symmetric	0.070	0.000	counterpart lower: met
FEVER	VitaminC	0.151	0.039	counterpart lower: met
ANLI-R1	ANLI-R3	0.128	0.086	counterpart lower: met
ANLI R1	R2, R3	$0.128 \to 0.049 \to 0.086$		monotone decline: *not met*

Table 22: Empirical coverage of the analytic diagnostic interval over 8 scenarios, including zero and near-zero Λ and sparse cells with four samples each. Coverage exceeds the nominal level in seven of eight scenarios, ranging from 0.963 to 0.997. The near-null two-cell case under-covers at 0.913 against nominal 0.95. This stress test does not establish general confidence coverage.

scenario	true Λ	per cell	empirical coverage	mean width
null, 4 types, n=100	0.000	25	0.993	0.326
null, 20 types, n=100	0.000	5	0.997	0.287
null, 50 types, n=200	0.000	4	0.993	0.190
near-null, 2 types, n=100	0.010	50	0.913	0.273
mild, 8 types, n=200	0.074	25	0.967	0.233
mild, 8 types, n=2000	0.074	250	0.963	0.076
strong, 8 types, n=200	0.237	25	0.970	0.201
strong, 8 types, n=2000	0.237	250	0.997	0.066
reference coverage target 0.95, 300 trials per scenario, worst 0.913				

detects severe sparsity but does not certify estimation accuracy. Table 24 reports both gates beside every measured value.

L Model-selection chronology and freeze audit

The fitting-disjoint claim in Appendix S.1 rests on a freeze whose content is verifiable. The feature map and θ were selected on the fitting sheet alone, disjoint from the evaluation sheet, and saved with a content hash, recorded with the model in the supplement's provenance package. The hash pins content, not chronology, and the commit history is explicit about the chronology it does not pin. The fit incorporated corrected annotations and was subsequently expanded to the full fitting sheet after evaluation annotation had begun. No evaluation row entered the fit, but the selection rule and the choice to fit on the full sheet were made with evaluation summaries visible. The saved model is verifiable, but model selection was not blind.

An earlier model, recorded there with its own hash, was fixed before the annotators returned their revised sheets and before the fit was widened to the full fitting sheet, so it embeds neither revision the reported model carries. Scored on the same fitting-disjoint split by the same code, it reaches corrected AUC 0.745 with gain 0.180 $[0.090, 0.279]$, against 0.743 and 0.178 $[0.088, 0.276]$ for the reported model. The two disagree substantially about the magnitudes of θ and agree exactly about the order θ induces over cells, so the correction-only AUC is identical at 0.714 and the observed gain persists for both saved models. Result 7 predicts that: only the induced order matters. This is a robustness check on the freeze and not a pre-registration claim. The evaluation sheets were already

Table 23: Survey of four benchmarks, including Climate-FEVER (Diggelmann et al., 2020), using published partial-input feature families. All arms passing the prespecified support and permutation-null gates are shown. The null is the 95th percentile of fifty label permutations, repeating the same selection procedure used for the observed labels. On GLUE-RTE's 277-row validation split, the null reaches 0.040, preventing a positive shortcut finding. The scientific-exaggeration corpus was excluded because its three-way severity labels had no published partial-input feature analogue. Every reported $\widehat{\Lambda}_{\mathrm{cf}}$ is a cross-fitted estimate, not a ceiling. Fitting coverage (Appendix F) is recorded per row in the artifact, and where it fails the value stands as a pairwise sensitivity statistic without the held-out-AUC interpretation.

benchmark	split	n	partial-input scheme	$\widehat{\Lambda}_{\mathrm{cf}}$	null p95	sensitivity range
AVeriTeC	dev	427	negation + length	0.082	0.048	[-0.003, 0.166]
AVeriTeC	train	2591	negation + length	0.042	0.021	[0.007, 0.076]
Climate-FEVER	test	907	negation + length	0.153	0.036	[0.091, 0.214]
SciTail	test	2126	PMI word + negation	0.082	0.020	[0.061, 0.102]
SciTail	train	23596	negation + length	0.025	0.007	[0.014, 0.037]
GLUE-RTE	validation	277	length	0.000	0.040	[-0.121, 0.080]
GLUE-RTE	train	2490	negation + length	0.012	0.013	[-0.021, 0.046]

Table 24: Λ measured, with Result 3a's empirical floor as a reporting diagnostic. Declined rows are those where the cross-fitted estimate falls below the floor: the lower end of the reported bracket is Result 3a's floor on the empirical Λ and the upper end the plug-in, which is upward-biased for the population Λ. The bracket is a reporting range on this sample and not a confidence interval. The gold-600 row uses nominal inverse-stratum calibration weights because the annotation sheets are enriched stratified designs. These are not exact inverse-probability weighting (IPW) under the realised comment-first sampler. Unweighted, the enrichment inflates the apparent asymmetry.

setting	n	cells	$\widehat{\Lambda}_{\mathrm{cf}}$	sensitivity range
FB15k-237 caged (head type)	325,620	94	0.112	[0.105, 0.119]
FB15k-237 within-type (head type)	325,621	93	0.019	[0.012, 0.026]
real errors (relation type)	100	4	0.279	[0.137, 0.420]
real errors (head,rel,tail types)	100	39	*declined*	bracket [0.296, 0.427]
gold 600, nominally weighted (relation type)	594	4	0.202	[0.129, 0.274]
gold 600, nominally weighted (head,rel,tail types)	594	113	*declined*	bracket [0.116, 0.323]

in the repository when that earlier model was fixed, both having entered in the same commit, so what it establishes is stability under refitting and not blindness. The same evaluation procedure was applied to each saved model.

M HELD-OUT-ACCOUNT PROBE: DESIGN AUDIT AND WITHDRAWAL

Prospectively specified, inferential analysis withdrawn. The design, success rule and reporting rule were committed with the sheet and before any verdict was seen. A design audit identified invalid inclusion weights, preventing the intended population inference.

The original annotation sample covers 47 of the corpus's 54 accounts. The remaining 7 accounts contain $1{,}478$ judgeable claims. The same three annotators labelled a sample of 145 claims under the original protocol, with verdict Fleiss $\kappa = 0.826$ compared with 0.823 on the evaluation sheet. Excluding one three-way tie as unresolved leaves 144 rows. The model was applied without refitting, with its hash checked against the saved evaluation model. This check verifies model content but does not establish blind model selection (Appendix L).

The build first allocated rows within strata and then enforced one claim per provenance comment across the entire sheet. Because comments can contribute claims to several strata, selections made in earlier strata changed the candidate pools in later ones. The stored n_h/N_h values are therefore not first-order inclusion probabilities. Neither the resulting nominally weighted AUCs nor a bootstrap

that treats those weights as fixed supports population inference. We retain the frozen design record, labels and original artifact, but withdraw the interval and make no held-out-account transfer claim.

N DISTRIBUTION SHIFT IN THE HELD-OUT SPLIT

Exploratory analysis. This split-level comparison was specified during a reproducibility audit after the outcomes were available. It is descriptive.

The fitting and evaluation sheets are hash-assigned halves of one pool under the same stratified design. Their nominally calibrated error rates are 36.6% and 25.6%, respectively, a difference of 11.0 percentage points with bootstrap sensitivity range $[3.3, 18.6]$. Standardizing the fitting split's within-relation rates to the evaluation relation mix accounts for 0.06 points, leaving 11.0. All four relations change in the same direction. The same annotators judged both sheets, with fitting annotation preceding evaluation annotation. This design cannot isolate annotation order as a cause of the rate difference.

The four relations retain the same error-rate ordering across splits. A random permutation reproduces that order with probability 0.042. After removing the common log-odds shift, the residual spread is 0.323. Stable cell ordering can preserve a cell-only AUC ranking (Result 7), but does not imply that refitting would leave θ unchanged.

O Λ AS A DATASHEET FIELD

A prospective map-selection protocol. State the suspected shortcut and the intended information restriction before inspecting the evaluation labels. Select a schema-defined map or a published partial-input feature family. Record its provenance, bin boundaries, missing-value handling and code version. Use the same frozen map in the compared benchmark conditions. If development labels guide a choice among maps, list the complete candidate family and selection rule, freeze the chosen map before an independent evaluation, and repeat the full selection inside each permutation used for a null comparison. Retain the results for unselected maps as descriptive analyses.

Report class counts per cell, empty and rare cells, the empirical value and support floor, and any population sensitivity analysis separately. A constant map is uninformative. A sparse map does not acquire a reliable population ceiling merely because its empirical maximum is large. Evaluate a simple prior on held-out labels when interpreting learned performance, and say how its order was obtained. Never choose the map by the detector whose score it is supposed to diagnose or describe a coarse map as bounding an unrestricted model. This protocol reduces researcher discretion. It does not identify a uniquely correct map. It is guidance for future benchmark audits, not a retrospective registration of our survey.

A benchmark report can state its cell map, empirical maximum AUC, permutation null and separate population sensitivity summary alongside trained baselines. The following example is the standalone tool's actual output. It keeps the exact finite-data identity distinct from the uncertain population interpretation, using only labels, cell keys and optional weights.

```
Type-asymmetry index (Lambda)
  cell key:               relation
  labelled items:         400 (116 positive)
  empirical Lambda:       0.2117 (exact on supplied rows)
  permutation null:       0.0840 (95th pct of 100 label shuffles)
  reading:                a cell-key-only score reaches at most AUC 0.7117 on
                          the supplied rows
  population sensitivity: cross-fit 0.2037
  verdict:                a cell-only shortcut is present. Report any
                          detector's score against a learning-free cell prior.
  method:                 empirical maximum-AUC identity, permutation null,
                          and cross-fit sensitivity
```

Three fields determine interpretation. **The cell key** specifies the information restriction to which the AUC ceiling applies. **The permutation null** measures finite-sample asymmetry under shuffled labels, using the same map-selection procedure as the observed statistic. Its gate is the 95th percentile of the shuffles. **The verdict** states the conclusion supported by the diagnostic and its reporting rules.

A large Λ motivates comparison with a cell-only prior. Reducing it by removing unsupported anomalies may substantially reduce the available evaluation pool: in the tested generator only 7.7% of candidates have attested signatures, and the replenishment cost is derived with Result 3a (Appendix E). This cost motivates changing the corruption procedure in that setting, without establishing that resampling is impossible at larger budgets.

P ALLOCATING THE ANNOTATION BUDGET

The error model is fitted from the mean feature vector of the errors, so the allocation should account for the number and weights of annotated errors. Under a stratified design an error found in stratum h carries weight N_h/n_h, and an allocation proxy is Kish's effective count

$$n_{\text{eff}} \approx \Big(\sum_h p_h N_h\Big)^2 \Big/ \sum_h p_h N_h^2/n_h.$$

This substitutes expected error counts into Kish's formula. It is an allocation heuristic, not an exact precision guarantee. Here h indexes strata, N_h is the stratum size in the pool, n_h the rows annotated in it, and p_h the probability that a claim there is wrong. The numerator does not depend on the allocation, so maximizing n_{eff} under a fixed total budget means minimizing $\sum_h p_h N_h^2/n_h$, and a Lagrange multiplier gives $n_h \propto N_h\sqrt{p_h}$. Explicitly, for total budget n and a positive allocation in each retained stratum,

$$\mathcal{L} = \sum_h \frac{p_h N_h^2}{n_h} + \mu\Big(\sum_h n_h - n\Big), \qquad \frac{\partial \mathcal{L}}{\partial n_h} = -\frac{p_h N_h^2}{n_h^2} + \mu = 0, \qquad n_h = \frac{n N_h\sqrt{p_h}}{\sum_j N_j\sqrt{p_j}}.$$

Here μ enforces the budget constraint. The per-stratum error rates are pilot estimates. Integer allocations, minimum coverage and strata with no observed pilot errors need practical safeguards beyond this continuous allocation formula.

Q COMPLETE MATCHED-NEGATIVE COMPARISONS

Figure 4 shows the crossed comparisons on FB15k-237 (Toutanova & Chen, 2015) and WN18RR (Dettmers et al., 2018). Table 26 gives the separate CCA intervention.

Inspection budgets. For CAGED, ADKGD and the corresponding priors, P@5% inspects $K = \lfloor 0.05 N_0 \rfloor$ items, where N_0 is the original graph size before anomaly injection. CCA uses $K = \lfloor 0.05 N \rfloor$, where N is the full ranked pool including anomalies. We retain these upstream conventions. Crossed comparisons and the CAGED–ADKGD reversal share a cutoff within each comparison, but the two architectures do not use identical fractions of their contaminated pools.

Synthetic labels and seeds. Injected anomalies receive synthetic error labels: a within-type replacement that happens to be a true but unrecorded fact is still labelled an error, so these rows measure agreement with benchmark labels rather than factual accuracy. CAGED/ADKGD reject candidates in the loaded graph. Their within-type generator tries 30 slot-attested replacements, then falls back to the published generator if none is accepted. The training runs did not log fallbacks, so the fraction realized in their training and injection streams is unknown. As a diagnostic replay, outside those random streams, we apply the acceptance rule once from every original graph triple at each of the four seeds. The replay falls back on 2.6% of draws on FB15k-237 and 0.0% on WN18RR. This estimates how often the rule falls back on these graphs. It is not a measured property of any original run. Accepted draws are slot-attested by construction. CCA rejects known triples and duplicate injected candidates, dropping infeasible source pairs from both evaluation columns. Absence from the graph does not establish factual falsity. CAGED and ADKGD inject their anomalies at load time from the same random state that seeds training. Each of their four seeds therefore varies both the injected-error draw and the optimization, and the two detectors receive identical anomalies at a given seed. CCA's five seeds vary training only, with the anomaly draw and pretraining checkpoint fixed by its protocol.

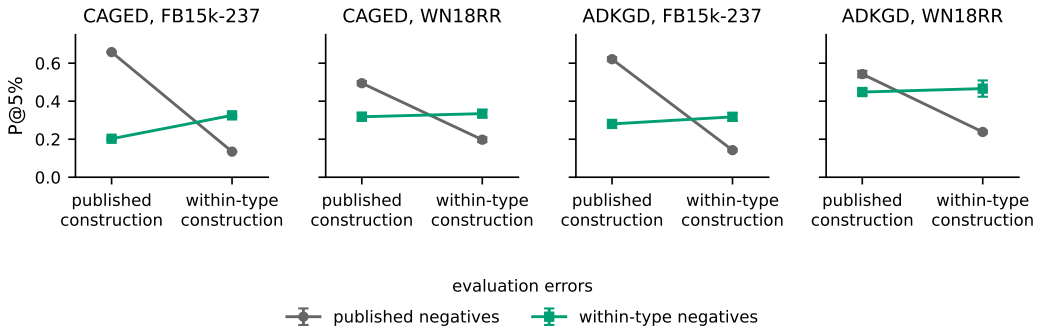

Figure 4: All four crossed comparisons inside two published detectors on two knowledge graphs. P@5%, mean ± standard deviation over four seeds. Only the corruption the detector trains against differs between the two lines. In every panel the lines cross: within-type negatives perform better on the within-type row and worse on the published row, so the preferred negatives are decided by which error process is audited.

Evaluation errors	Published negatives	Within-type negatives
Published construction	**0.6583 ± 0.0031**	0.2025 ± 0.0064
Within-type construction	0.1346 ± 0.0014	**0.3255 ± 0.0056**

Table 25: CAGED on FB15k-237, P@5%, mean ± standard deviation over four seeds. Rows are the errors audited, columns the negatives trained against. Compare within a row. The diagonal compares two different benchmarks. The top-left cell is the published configuration, reproduced exactly by the patch. Bold marks the better training condition in each row.

Evaluation errors	Control negatives	Matched negatives
Published construction	0.4753 ± 0.0054	0.4548 ± 0.0037
Within-type construction	0.4556 ± 0.0063	0.4683 ± 0.0038
interaction per seed: +0.0500 +0.0261 +0.0346 +0.0308 +0.0245		

Table 26: The prespecified CCA 2 × 2 on FB15k-237, P@5%, mean ± standard deviation over five seeds. Axes are as in Table 25. CCA is architecturally independent of the CAGED lineage. The protocol, its input checksums and the directional prediction were frozen before any primary-metric outcome existed. Relative to the control, the matched arm changes P@5% by -0.020 on the published row and $+0.013$ on the within-type row, and the interaction is the second change minus the first. All seed-level interactions are reported under the prespecified reporting rule.

CCA's published training negatives. Among the replacement anomalies of CCA's benchmark, 62.1% are type-preserving: the substituted entity is already observed in that argument slot for that relation. CCA's training negatives are built differently, by permuting arguments within a shuffled minibatch, so the substitute comes from a uniformly drawn other triple without conditioning on the relation. An exact simulation of one pass of that permutation over the training triples, with nothing trained, tags each argument-replacement negative by the same rule on the same graph. Only 7.7% are type-preserving, and whole-triple swaps, which carry no type verdict, are excluded. CCA thus evaluates against a largely type-preserving construction but trains against a largely type-violating one, the mismatch its crossed intervention tests.

Specification of the CCA intervention. CCA has no corruption generator to replace, so the matched arm importance-reweights the in-batch permutation negatives it already builds. For a tail-replacement negative (h, r, t') whose donor tail t' comes from the shuffled batch, the weight is $w = \Pr(t' \mid \text{tail slot}, r) / \Pr(t' \mid \text{tail slot})$, and symmetrically for a head replacement. Both slot

tables are counted over the public graph's own triples, the transductive pool CCA trains and scores on. An unattested donor is assigned $w = 0$, so the type-preserving restriction follows from the weighting rather than being imposed. A permuted triple that coincides with a real graph edge is also assigned $w = 0$. This collision rejection removes 2.1% of in-batch negatives at batch size 512. Weights are self-normalized within the batch, so the loss keeps its scale and the two arms stay comparable. The control arm uses the same draw count, so only the weighting differs. At batch size 512 with 16 permutation draws per slot, reweighting moves the negatives from 9.9% type-preserving to 100.0% of the weighted mass, at 235.7 effective negatives per batch (Kish, 46.0% of the batch). The measured effect is that joint intervention, not the weighting alone.

The weighting has two separate implementations, one standalone and one integrated into CCA, and their weight tables agree exactly on 2000 sampled triples. The protocol, the five input checksums, the build seed 20260830, the four cells, the primary metric (Precision@5% at epoch 20), the directional prediction and the reporting rule were frozen before any primary-metric outcome existed. The protocol record in the supplement's provenance package holds these checksums and the content checksums of the selected triples and both paired evaluation anomaly sets.

For the control, both injected anomalies and training negatives use the published configuration. In this setting, the patch reproduces CAGED's committed four-seed aggregate to full float precision on both graphs. The crossed runs retain the remaining detector settings while changing the designated corruptions. ADKGD's unmodified runs were not re-aggregated, so equivalent control reproduction is unverified for that system. The artifacts record this distinction. All four crossed tables show the same direction: within-type training negatives improve P@5% on the within-type row and reduce it on the published row. On CAGED's within-type rows, the improvement factors range from $1.69\times$ to $2.42\times$.

Evaluation errors	Published negatives	Within-type negatives
Published construction	**0.4947 ± 0.0115**	0.3182 ± 0.0209
Within-type construction	0.1974 ± 0.0127	**0.3346 ± 0.0192**

Table 27: CAGED on WN18RR, P@5%, four seeds.

Evaluation errors	Published negatives	Within-type negatives
Published construction	**0.6210 ± 0.0108**	0.2804 ± 0.0064
Within-type construction	0.1424 ± 0.0076	**0.3176 ± 0.0209**

Table 28: ADKGD on FB15k-237, P@5%, four seeds.

Evaluation errors	Published negatives	Within-type negatives
Published construction	**0.5427 ± 0.0170**	0.4482 ± 0.0169
Within-type construction	0.2381 ± 0.0078	**0.4663 ± 0.0428**

Table 29: ADKGD on WN18RR, P@5%, four seeds.

R Comparison with prior controls

Figure 5 compares each correction with its prior control on the same evaluation sample.

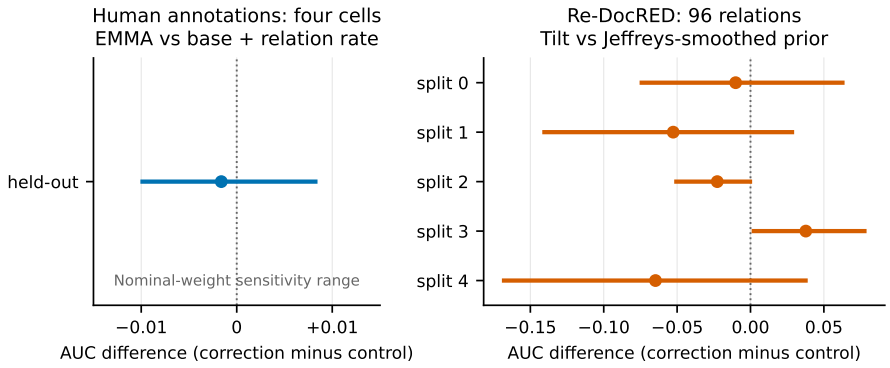

Figure 5: Comparison with prior controls. Dots are differences of the reported point AUCs, not bootstrap means. Left: EMMA minus base plus relation rate on the fitting-disjoint evaluation sheet. The nominal-weight sensitivity range spans zero. Right: tilt minus a Jeffreys-smoothed relation prior in each Re-DocRED split, with 95% document-bootstrap percentile intervals. The splits overlap and are not independent replications. The panels use different horizontal scales and uncertainty conventions. Neither supports predictive superiority over its control.

S THE FITTING-DISJOINT EVALUATION RESULT, PER SCORE

Figure 6 summarizes the correction, whose score-level results appear in Table 30.

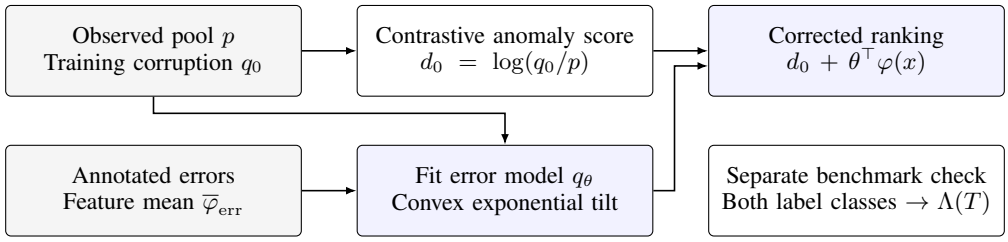

Figure 6: EMMA's information flow. The score identity holds for a contrastive log-density ratio. Adding the correction to other scores requires empirical validation. The fit uses annotated errors and samples from q_0. The separate Λ diagnostic uses both class marginals on a stated cell map T. It bounds scores on that map, not the unrestricted detector.

S.1 A DERIVED CORRECTION AND ITS LIMITS

The natural response to §3 is to move the corruption toward labelled errors. For a feature map φ, the exponential tilt $q_\theta \propto q_0 \exp(\theta^\top \varphi)$ is fitted by a convex moment-matching objective in which only annotated errors enter the fitting term (Result 6). It changes the contrastive score by

$$d_\theta = \log(q_\theta/p) = d_0 + \theta^\top \varphi - \log Z(\theta), \qquad d_0 = \log(q_0/p), \tag{33}$$

where $Z(\theta)$ is the normalizer, so ranking only requires adding $\theta^\top \varphi(x)$ to the base log-ratio (Appendix E). The identity is exact for a base log-density-ratio score. The structural auditor below is not calibrated to that scale, so the implemented sum is an empirical score combination (see the scale sensitivity below). On 300 fitting-disjoint evaluation rows (model selection was not blind, Appendix L), the correction raises weighted AUC from 0.564 to 0.743, but adding the raw relation rate gives 0.744, with paired difference -0.002 and sensitivity range $[-0.010, 0.008]$. A smoothed prior also outperforms the tilt on Re-DocRED (Table 30 and Appendices T and R). The correction is a principled response to the theory, not an improvement over label-matched priors, and the population margin bounds do not certify it (Appendices W and V).

Table 30: The frozen error model on 300 fitting-disjoint annotated rows, joined to the structural reliability score already computed on the same graph. Nominal calibration weighting is used for the enriched design. Ranges are intended-design sensitivity summaries. The gain is a paired bootstrap over 5000 draws.

score	calibrated AUC	sensitivity range
base auditor $1 - r$ (structural reliability alone)	0.564	[0.483, 0.643]
frozen correction alone ($\theta^\top \varphi$, relation type)	0.714	[0.643, 0.792]
EMMA-corrected $(1 - r + \theta^\top \varphi)$	0.743	[0.668, 0.817]
gain over the base auditor (bootstrap mean)	+0.179	[0.088, 0.276]

The extractors. Both extractors are local Ollama models processing the same comments through the same prompt and the same fixed output schema. Both decode greedily at temperature 0 with a 1500-token cap and no thinking trace, so the model is the only difference between them. The claims audited here, and the gold sheets drawn from them, come from Qwen3-32B. The second extractor of Appendix V is Qwen3-8B, re-run over the same comments. Parses that fail the schema are dropped rather than repaired.

The base auditor. The structural reliability score is a weighted mean of the components available for an edge, $r = \sum_i w_i c_i / \sum_i w_i$ with each $c_i \in [0, 1]$, so $r \in [0, 1]$ and the anomaly score is $1 - r$. The label-free components are balance consistency $(w = 0.25)$, co-occurrence coherence (0.20) and sign margin (0.15). The simple priors are extractor confidence (0.15), schema validity (0.10) and evidence-span containment (0.15). Counts enter through $x \mapsto x/(x + m)$, with m the median positive count, so an unavailable component is dropped from both sums rather than imputed. r uses no error labels. The components are defined as follows. Balance consistency is the fraction of the signed triangles an edge feeds that are balanced, and co-occurrence coherence is the saturated co-occurrence weight of its endpoints. Sign margin is the mean agree-minus-disagree margin of the co-stance ties it feeds, and extractor confidence is the extractor's own confidence. Schema validity records whether both endpoints carry types the edge permits, and evidence-span containment whether the quoted evidence span occurs in the source comment. The first three are label-free and the last three are simple priors. A relation edge has no balance or margin component, since relations do not enter the signed graph.

Two sheets of 300 rows each were annotated independently by three annotators, at verdict Fleiss κ 0.868 and 0.823. Six unresolved fitting labels were excluded, leaving 594 resolved rows including all 300 evaluation rows, and no evaluation row entered the fit or the feature map. In this instance q_0 is the nominally weighted fitting sheet itself: all 300 fitting rows enter the normalizer with their labels unused. With relation indicators, differences in θ therefore equal log ratios of weighted relation error rates, up to 3.5% ridge shrinkage of their spread. The correction is thus a log relation-rate prior by construction, which is why the rate controls of Appendix T match it. The correction alone reaches 0.714 against 0.743 for the combined score. These are separate measurements rather than an additive decomposition. Dropping the nominal weights entirely gives $0.571 \to 0.719$, so the gain does not depend on the weighting whose design defect Appendix M records.

Scale sensitivity of the combination. The base auditor is not calibrated to the log-density-ratio scale. Adding $\theta^\top \varphi$ to it is therefore a score combination rather than the exact identity of (33), and the combined ranking depends on the base's scale in a way its own AUC does not. Rescaling the base by k leaves its AUC unchanged and moves the gain over $[0.178, 0.180]$ for $k \in [0.01, 2]$, reaching 0.116 at $k = 10$. The gain persists over the tested scales, but its magnitude depends on scale: at a large enough k the base dominates and the correction survives only by breaking its ties. A discrete base with many ties can retain a gain for that reason alone.

T LANGUAGE-MODEL AND NLI BASELINES

A language-model judge supplies a stronger base score. On the same frozen 300-row evaluation split, Qwen3-32B (Yang et al., 2025) with 3-vote self-consistency reaches weighted AUC 0.720 $[0.650, 0.786]$. Adding the frozen correction raises it to 0.774. The paired gain has sensitivity range

$[0.017, 0.093]$. That comparison establishes a gain over the judge alone, not an advantage over other corrections using the same fitting labels.

Relation-prior controls on the same judge. An exploratory follow-up uses all three cached votes for each evaluation row, without new model calls or row exclusions. Every prior uses only the same 294 resolved fitting rows. For relation r, let

$$W_r = \sum_{i \in \text{fit}: r_i = r} w_i, \qquad E_r = \sum_{i \in \text{fit}: r_i = r} w_i y_i, \qquad \widehat{p}_{r,\alpha} = \frac{E_r + \alpha}{W_r + 2\alpha}.$$

Here w_i is the nominal calibration weight, y_i the error indicator, and r_i the relation. We use $\alpha = 0$ and $\alpha = 1/2$, retaining the weight scale of the existing same-budget control. The half-count is smoothing of weighted counts, not a posterior uncertainty statement. The fixed controls add either $\widehat{p}_{r,\alpha}$ or its log-odds to the judge's fraction-of-wrong-votes score. Probabilities are clipped to $[10^{-6}, 1 - 10^{-6}]$ before taking log-odds. All additive coefficients are one in Table 31, as for the reported EMMA score.

score	weighted AUC	Qwen+EMMA − row	sensitivity range
Qwen judge alone (fraction of wrong votes)	0.7202	–	–
Qwen + **EMMA** ($\theta^\top \varphi$)	0.7737	–	–
Qwen + fitting-set relation rate $\widehat{p}_{r,0}$	0.7742	-0.0005	[-0.0094, 0.0084]
Qwen + smoothed relation rate $\widehat{p}_{r,1/2}$	0.7742	-0.0005	[-0.0094, 0.0084]
Qwen + relation log-odds, $\alpha = 0$	0.7720	0.0017	[-0.0179, 0.0204]
Qwen + relation log-odds, $\alpha = 1/2$	0.7720	0.0017	[-0.0179, 0.0204]

Table 31: Cached Qwen judge and fitting-set prior controls. AUCs and differences are point estimates. The third column is Qwen-plus-EMMA minus the row's score. The final column is its paired 5,000-draw intended-design sensitivity range conditional on the fitted scores. It is not confidence coverage for the realized sampler. All four controls are reported.

The raw-rate and smoothed-rate controls both reach 0.774, versus 0.774 for EMMA. Their difference is -0.0005 $[-0.0094, 0.0084]$: these corrections are not separated. Both log-odds controls are also reported. The Qwen result therefore shows that fitting-set relation information can help a strong judge. It does not distinguish the exponential tilt from the tested relation-prior alternatives. Table 32 varies the coefficient on the correction to show how much of that comparison depends on its scale.

correction	$k = 0.25$	$k = 0.5$	$k = 1.0$	$k = 2.0$	$k = 4.0$
EMMA ($\theta^\top \varphi$)	0.773	0.774	0.774	0.770	0.762
fitting-set relation rate $\widehat{p}_{r,0}$	0.773	0.773	0.774	0.776	0.772
smoothed relation rate $\widehat{p}_{r,1/2}$	0.773	0.773	0.774	0.776	0.772
relation log-odds, $\alpha = 0$	0.775	0.775	0.772	0.762	0.762
relation log-odds, $\alpha = 1/2$	0.775	0.775	0.772	0.762	0.762

Table 32: Descriptive scale sensitivity: AUC of the Qwen score plus the column's coefficient times the row's correction. No coefficient was selected on these evaluation results. This grid is not a fitted fusion model or an additional confirmatory test.

The control comparison was motivated after the original evaluation had been examined, so it is exploratory. A learned joint Qwen-and-relation model would require Qwen scores on the fitting rows and is not tested here. The judge received the same underlying evidence as the extractor. It is not an independent source of truth. Joint weighted regression coefficients are descriptive: treating nominal weights as frequency weights would mistake weighted corpus mass for replicated observations. A zero-shot multilingual NLI comparator, mDeBERTa-v3-base fine-tuned on XNLI and multilingual NLI data,[1] reaches 0.653. The same frozen correction lifts it by 0.068 $[0.020, 0.117]$, and the corrected structural auditor exceeds it by 0.090 $[0.015, 0.164]$. That NLI comparison does not supply the missing prior control for its own base score.

[1] Checkpoint training data and evaluation are documented in the original model card.

U SCHEMA-TYPE PRIOR CONTROLS

Figure 7 compares schema-type priors with the published detectors under each anomaly construction.

Definition of the prior. Give each entity the Freebase type it most often heads, the first two components of the relation path (`/people/person` for `/people/person/nationality`, the subject's path for a mediated relation), with ties broken lexicographically, and write $\tau(e)$ for that type. The type map is derived either from the contaminated graph the detectors train on (the equal-information prior of §5) or from the original graph *before anomaly injection*, which is additional benchmark information. Counting support over the contaminated graph available to the auditor, the score of a triple (h, r, t) is

$$-\log\big(1 + [N(\tau(h), r) - 1]_+\big) - \log\big(1 + [N(r, \tau(t)) - 1]_+\big),$$

higher being more anomalous, where $N(\tau, r)$ counts triples whose head has type τ under relation r and $N(r, \tau)$ the same for tails. Subtracting one removes the scored triple's support for itself. No detector is trained, and injected-error labels are not used by the scoring function. WN18RR's relation names carry no schema, so only the argument-identity variant applies there.

Information access. CAGED and ADKGD train and score on the contaminated pool, use no injected-error labels and load no pretrained embeddings. Their negative generators do reject candidates that occur in the pre-injection graph, and the within-type generator of Appendix Q also draws replacements from pre-injection slot pools. The contaminated-graph type map, a post-hoc control, gives the prior no information beyond the contaminated pool, with the score and the cutoff unchanged.

Paired comparison on the detectors' pools. CAGED and ADKGD inject their anomalies when their loader reads the graph, from the random state their training scripts seed. We rebuild each seed's pool by running the detectors' own loader under that seeding, with the pinned input files, and check that CAGED and ADKGD produce the same multiset of (triple, label) pairs at every seed, and that our port of the construction produces it too. All four seeds agree triple for triple, so the prior is scored on the pools the detectors were evaluated on and each seed can be paired with the detectors' logged P@5% on the published configuration. The detectors' four-seed means recovered this way are the published 0.658 ± 0.003 and 0.621 ± 0.011 exactly. The equal-information prior reaches 0.680 ± 0.002 (AUC 0.943). Paired by seed, it exceeds CAGED by 0.022 on average (at least 0.018) and ADKGD by 0.059 (at least 0.041), in every seed. With the pre-injection map it reaches 0.707 ± 0.002. On these pools the prior's lead over both detectors therefore does not require the pre-injection map. A sign consistent over four paired seeds is not a population-level interval.

Increasing the fraction of injected anomalies drawn from the published construction from 0 to 1 raises a schema-type prior's AUC monotonically from 0.521 to 0.955. On the same pools, Λ on the coarser head-type map rises from 0.019 to 0.116. Neither calculation requires training. On WN18RR, by contrast, CAGED reaches 0.495 ± 0.011, exceeding the argument-identity prior's 0.200 on five pools drawn by our port of the construction. Thus prior superiority is benchmark-specific. Matched training negatives still improve auditing on WN18RR (§4.1), so that intervention's benefit does not require a prior to outperform the detector.

V ADDITIONAL TRANSFER AND HUMAN-LABEL ANALYSES

On Re-DocRED the method refits and reaches AUC 0.859 against an in-sample reference of 0.885, over 148 held-out errors.

Across languages. Models fitted separately on English-only and non-English-only fitting subsets give mean gains 0.188 across languages and 0.179 within languages, with all sensitivity ranges excluding zero. The four-cell relation map contains no language-specific features, and its fitted θ closely tracks relation error rates (Spearman 1.00). These comparisons assess transfer of this coarse correction. They do not establish noninferiority to within-language fitting without a prespecified margin.

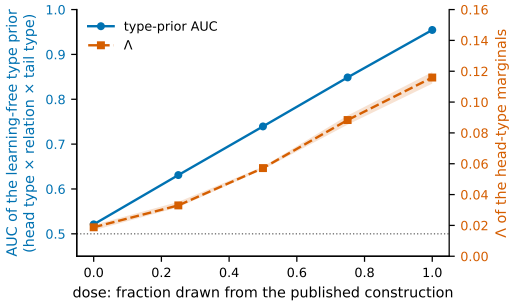

Figure 7: Dose-response. Sweeping the fraction of injected anomalies drawn from the published construction moves a learning-free schema-type prior's AUC monotonically, with Λ measured on the same pools beside it. Nothing is trained. The prior's AUC exceeds $\frac{1}{2} + \Lambda$ at every dose. This is consistent with Result 4, which bounds a score only by the index of its own cell map (§5): the score keys on both head-type–relation and relation–tail-type features, while the Λ plotted beside it indexes head-type marginals alone.

Transfer of cell ordering across extractors. The frozen θ reaches AUC 0.882 on the second extractor's claims, which is exactly what a prior fitted on those labels directly attains. Over a four-cell map every score is constant within a cell, so the AUC depends only on the order the score puts the cells in, and any two scores agreeing on that order tie by construction. The substantive result is the order itself: *causes* $>$ *role of* $>$ *part of* $>$ *associated with* on both extractors, while the rates underneath change substantially, *associated with* from 0.229 to 0.011 and *part of* likewise. This measures stability of the correction's cell ordering across an extractor on which it was not fitted. It does not by itself verify the density-ratio assumptions of Result 7. Six rows the majority could not settle, four split three ways and two with the channel tied, were resolved by the three annotators together with the schema rule in front of them, after the independent round had closed. All 150 rows carry a verdict.

Errors in human-authored entity labels. CoNLL++ tests whether type-dependent error rates also occur in human-authored labels. The CoNLL-2003 test set (Tjong Kim Sang & De Meulder, 2003) was re-annotated by hand in CoNLL++ (Wang et al., 2019b), which corrected label mistakes in 5.39% of its sentences. Diffing the two by entity span gives 142 real errors over 5648 asserted entities, identified by human revision of human-authored labels.

The error rates vary substantially by entity type. 6.6% of miscellaneous entities are wrong against 0.9% of person, a sevenfold spread over four cells, and the estimated cross-fitted $\widehat{\Lambda}_{\mathrm{cf}} = 0.183$ $[0.119, 0.247]$ puts an out-of-fold cell prior at AUC 0.683 on this map and label contrast. Fitting coverage passes in every direction here, which is what licenses that interpretation: the estimator emits this field only when the condition of Appendix F holds, and reports the raw statistic alone otherwise. It stands beside our own corpus at 0.202 and Re-DocRED at 0.390, showing type-dependent rates in three independently annotated error processes, including human-authored labels.

The rates also transfer. A sentence-level split keeps entities from the same annotation context together. Rates fitted on one half reach AUC 0.717 on the other against an in-sample reference of 0.718 that uses the held-out labels directly. The gap is smaller than the corresponding Re-DocRED gap.

The implementation verifies identical token sequences between releases before matching entities by span. It also verifies zero differences on the training and development splits, which CoNLL++ did not revise.

Re-DocRED supplies real errors with no new annotation: the triples human annotators removed from DocRED. Its relation marginals give $\widehat{\Lambda}_{\mathrm{cf}} = 0.390$, the highest in Figure 2, and a learning-free relation prior reaches 0.881 AUC there. Two limitations apply: Re-DocRED is *predominantly a completeness revision* (58787 triples added against 275 removed), so it supplies real false positives at a low rate, and 69.5% of its errors lie in 2 of 96 relations.

W RESIDUAL ERROR AGAINST REVIEW BUDGET

Reviewing the top 30% of nominally weighted corpus mass and repairing what is found leaves 11.3% of the corpus wrong under the corrected ordering against 16.0% under the base. This is a retrospective sensitivity calculation and supplies no accept-set guarantee.

Residual-error calculation. Group equal scores and order groups from highest to lowest. Let W_g and E_g be the total nominal weight and error weight in group g, and put $W = \sum_g W_g$. For review budget $b \in [0, 1]$, the fraction reviewed in group g and residual error are

$$a_g(b) = \min\left(1, \max\left(0, \frac{bW - \sum_{j<g} W_j}{W_g}\right)\right), \qquad R_s(b) = \frac{\sum_g E_g[1 - a_g(b)]}{W}.$$

Here s denotes the ranking. Fractional review at a tied score spends exactly the same budget under each ranking and averages within that tied group. The denominator is the original total weight, not the weight of unreviewed items. The calculation assumes perfect repair of errors found. The random comparator is one seeded random ordering, and the oracle orders errors first. The gap column in Table 33 reports the mean paired-bootstrap difference rather than the difference of the adjacent point estimates.

Interpretation and inferential limits. The curve is descriptive. It says what a given review budget would leave wrong if the reviewed errors were repaired perfectly, under the ordering each score induces. The sensitivity analysis uses the intended stratified design. The sample is stratified and enriched, so the audit-curve analysis redraws rows inside their own stratum at its own size rather than uniformly. Correlated claims from one comment would require an additional cluster level, and here there are none: all 300 eval rows come from distinct comments, as checked by the implementation. The earlier certificate is withdrawn because its threshold search selected and tested on the same sample. We also withdraw an earlier numerical sample-size projection: it held the number of searched thresholds fixed as the sample grew and plugged in a residual from the invalidly selected threshold. No valid replacement certificate or label requirement is claimed here.

At each reported budget below full review in Table 33, the pointwise 95% sensitivity interval excludes zero. At full review both residuals are zero. The bootstrap constructs no simultaneous band, so nothing is claimed about all budgets at once. It is a paired Rao–Wu rescaled bootstrap under the stratified design the sheet was *intended* to have. Selection enforced one claim per comment across strata in iteration order, so the stored weights are nominal stratum fractions rather than realised inclusion probabilities. These are therefore sensitivity intervals, not exact design inference. The budget axis inherits the same caveat, being calibrated rather than design-unbiased corpus mass.

budget	random	base	corrected	oracle	base − corrected
5%	0.246	0.241	0.220	0.206	+0.020 [0.005, 0.035]*
10%	0.233	0.216	0.193	0.156	+0.025 [0.003, 0.045]*
20%	0.213	0.189	0.156	0.056	+0.032 [0.002, 0.061]*
30%	0.188	0.160	0.113	0.000	+0.048 [0.012, 0.081]*
40%	0.164	0.131	0.081	0.000	+0.051 [0.013, 0.086]*
50%	0.136	0.102	0.055	0.000	+0.048 [0.016, 0.079]*
75%	0.060	0.060	0.023	0.000	+0.036 [0.013, 0.057]*
100%	0.000	0.000	0.000	0.000	+0.000 [0.000, 0.000]

Table 33: Share of the corpus still wrong after reviewing the top b of corpus mass under each ranking and fixing what is found. Budget is corpus mass rather than rows, since the sheet is an enriched stratified sample and each reviewed row carries its nominal stratum weight. The oracle reviews true errors first and is the best any ranking could do. The random comparator uses a seeded random ordering. Starred gaps exclude zero over a paired bootstrap of 5000 draws. The model assumes review is perfect and introduces no new errors.

X RE-DOCRED CORRUPTION COMPARISON: DESIGN AND AMENDMENT

The design, the single confirmatory statistic and the reporting rules were committed before the first run. One corrective amendment followed the first result. It fixed the tie convention and added the coupled interval described below. No split, draw or fitted model changed, and all ten seeds were rerun.

Data. DocRED assertions, keyed by entity name and de-duplicated because the two releases index their entity sets independently. $Y = 1$ is an assertion the Re-DocRED revision removed, $Y = 0$ one it retained. The revision's added triples are excluded: they are omissions, a different error mechanism, and mixing them would change the task.

Design. Documents are partitioned before their triples are assigned to fitting and evaluation, with document disjointness checked by the implementation. One contrastive auditor, identical in every arm in features, optimizer, iterations, ridge and accepted-negative count, separates the fitting half's assertions from negatives drawn from that arm's corruption. Its anomaly score is the negated logit. Only the corruption differs. q_{matched} is q_{random} reweighted by the smoothed cell marginal of the fitting half's removed triples, so its support equals q_{random}'s and it uses no evaluation label. We verified this by recomputing its weights from the fitting half alone and confirming they differ from the full-pool version.

Auditor specification. Relation, head-entity type and tail-entity type are one-hot encoded using a vocabulary fitted on the training pool. The scalar features are head and tail degree, head and tail mention count, evidence count and a same-sentence indicator. Each count is transformed as $\log(1+x)$. The design is standardized on the training pool, then augmented with an intercept. Fitting is ridge-penalized logistic regression by Newton steps: ridge 0.01, at most 200 iterations, stopping at a maximum step of 10^{-10} and logits clipped to ± 30. It separates 4000 positives drawn from the fitting half from 4000 negatives drawn from that arm's corruption, over 10 document-disjoint splits, with score ties taken at 12 decimals. The matched corruption's cell is the triple of relation, head-entity type and tail-entity type, and its weight on a candidate in cell c is $(n_c + \alpha)/[(N + \alpha C)\, r_c]$, normalized to a distribution, where n_c counts the fitting half's removed triples in c, N their total, C the number of occupied candidate cells, r_c the candidate count in c, and $\alpha = 0.5$ is the Jeffreys smoothing constant. An unattested cell keeps positive weight through α, so the support is q_{random}'s. The reproduction checks rerun this procedure with the recorded settings.

Analysis. The confirmatory statistic is the seed-averaged paired difference $\mathbb{E}_s[\mathrm{AUC}_s(q_{\mathrm{matched}}) - \mathrm{AUC}_s(q_{\mathrm{random}})]$, with each of 5000 bootstrap replicates resampling that seed's own evaluation documents, giving the prespecified interval $[0.282, 0.328]$. Because the ten splits share documents, the interval reported in Table 1, $[0.253, 0.357]$, instead couples a document's multiplicity across every split in which it appears, and is wider. AUC was fixed as the primary metric in advance because the retained-to-removed ratio is a property of the corpus rather than of a deployment. An integrity check verifies the construction and is deliberately unable to compute the primary metric, so that confirming the design is sound cannot reveal how it performs.

Scope. The permutation and smoothing-only baselines were added after the primary outcome, as falsification checks rather than confirmatory tests, and are reported as such. They do not alter the prespecified statistic. They assess the contribution of fitting-label information relative to these controls. The pseudo-count $\alpha = 0.5$ per cell carries 0.89 of the matched weight, and pseudo-counts alone, using no labels, give a gain of 0.148. Permuting the fitting half's labels gives a gain of 0.136 on average over 20 permutations, at most 0.144, so the gain the labels add beyond permutation is 0.168. The relation-prior control belongs to the same post-hoc group: each relation's smoothed removal rate in the fitting half reaches 0.879 on the same splits, ahead of the matched arm by 0.053.

Y A SECOND HUMAN ERROR PROCESS: THE CONLL++ NULL RESULT

CoNLL++ (Wang et al., 2019b) provides a second human error process in named-entity annotation. The matched-versus-random contrast is unresolved on this task.

Design. Identical in form to Appendix X, and specified the same way, before the first run. One logistic auditor, the same feature map, ridge and accepted-negative budget in every arm, on document-disjoint halves over 10 seeds of the 231 CoNLL-2003 test documents CoNLL++ revised. $Y = 1$ is

one of the 142 entity assertions the revision deleted or retyped, among 5648 asserted. $Y = 0$ is one it kept. Only the training corruption differs across arms. The same diff is analysed from the Λ side in Appendix V. Here it supplies the evaluation errors rather than a measurement.

Table 34: Held-out AUC by training corruption on CoNLL++, seed-averaged over 10 document-disjoint splits. The matched-versus-random difference is the prespecified confirmatory statistic. The type arm is descriptive and does not enter the confirmatory decision.

Training corruption	Mean held-out AUC	Role
q_{random}	0.729	Primary comparator
q_{matched}	0.750	Primary matched arm
q_{type}	0.655	Descriptive only

Table 34 gives the three arms. The primary paired difference is 0.0212, with a coupled-document 95% bootstrap interval of $[-0.0145, 0.0572]$ over 5000 replicates, and 7 of 10 seeds positive. The interval contains zero and does not establish a positive matched-versus-random effect.

Interpretation. The interval does not distinguish a small effect from insufficient precision. CoNLL++ provides 142 errors, with approximately half available for fitting in each split, fewer than in Re-DocRED. This sample-size difference does not establish the cause of the unresolved contrast. The intervals also condition on fitted scores and splits, omitting fitting uncertainty. Result 1 permits corruption-dependent rankings without requiring a detectable finite-model effect on every task. This study tests a second human error process within structured extraction, not a second crossed 2×2 interaction.

Protocol amendment before outcome evaluation. The first fit under the frozen protocol (seed 0, q_{random}) reached its iteration cap without converging. No arm returned scores and no AUC was computed. The producer-generated failure artifact is preserved in the repository history. A dated protocol addition then raised the maximum iterations to 2000 (committed after that attempt and *before any AUC existed*) and all 10 seeds were restarted from scratch. Every scientific setting and tolerance was unchanged, and all fits then converged. No setting was altered and no seed was added after the AUCs were seen.

Z IMAGE INTERVENTION: DESIGN AND PROTOCOL AMENDMENT

Design. The observed pool is $P = 0.90\,C + 0.05\,E_N + 0.05\,E_D$, with C CIFAR-10, E_N CIFAR-100 (Krizhevsky, 2009) and E_D SVHN cropped digits (Netzer et al., 2011). The component labels construct the pool and are withheld from every fit. Both error families occur unlabelled inside P, so this is not an unseen-family experiment. Two auxiliaries are used: Tiny ImageNet-200 objects (Le & Yang, 2015), reduced by re-applying the published CIFAR-relevance exclusion so no retained class is a CIFAR-10/100 synset or one of their ancestors or descendants, and EMNIST Letters (Cohen et al., 2017). Two backbones, ResNet-18 (He et al., 2016) and WRN-28-10, a wide residual network (Zagoruyko & Komodakis, 2016), are trained from scratch on 20000 CIFAR-10 images and frozen, so no third-party checkpoint can overlap a source. The head is a linear logistic classifier on those frozen features, auxiliary against pool, standardised on the fitting pool alone and identical across arms.

Difficulty matching. Each auxiliary is diluted into the pool, $Q_{j,\lambda} = (1 - \lambda)P + \lambda Q_j$, and λ is chosen so the two arms agree within the stated tolerance in pool-versus-auxiliary discrimination AUC. Matching uses only pool and auxiliary data. No evaluation record is accessed. Selection and confirmation use disjoint data, and the confirmation criterion (the paired 90% interval wholly inside ± 0.02, with each arm's one-sided 95% lower bound above 0.60) was fixed in advance. Both backbones passed, with paired 90% intervals $[-0.0048, 0.0047]$ and $[0.0019, 0.0078]$, both wholly inside the ± 0.02 tolerance fixed in advance.

Development-informed amendment. The originally specified grid, $\lambda \in \{0.25, 0.50, 0.75, 1.00\}$, produced no pair qualifying on *both* backbones, which the prespecified rule required. ResNet-18 had no eligible pair at all, its closest differing by 0.0239 against the 0.02 tolerance. WRN-28-10 had

one, at $\lambda_N = \lambda_L = 0.25$, which ResNet-18 did not share. The original matching criterion therefore failed. For any fixed score s, write $\mathrm{AUC}(s, Q, P)$ with auxiliary draws positive and pool draws negative. With $Q_\lambda = (1-\lambda)P + \lambda Q$,

$$\mathrm{AUC}(s, Q_\lambda, P) = \tfrac{1}{2} + \lambda \left[\mathrm{AUC}(s, Q, P) - \tfrac{1}{2}\right].$$

Thus arms with different separation ceilings are matchable inside their overlapping ranges, and the failure was grid resolution rather than a limit on dilution. The amended protocol, written before target access, refined the grid to a step of 0.05, and matching was asserted per backbone rather than through a single shared mixture, because the two representations require different ratios (0.798 and 0.905). Matched difficulty is a property of the representation, not of the sources alone. This amendment relaxed the original requirement that one mixture serve both backbones. It used observed development matching curves before target access and was not part of the initial specification. The tolerance, the floor, the confirmation procedure and every strong criterion were carried over unchanged. Appendix Z.1 reports the complete refined grid.

Full results. Targets were opened once, after confirmation passed, and Table 35 reports every target effect.

Table 35: Target effects with 95% intervals and one-sided 98.75% Bonferroni lower bounds, over 20000 bootstrap draws resampling each target component separately with the clean set shared across both comparisons.

Backbone	Effect	Point	95% CI	Bonferroni lower
ResNet-18	D_N	$+0.301$	$[0.292, 0.311]$	$+0.290$
ResNet-18	D_D	$+0.044$	$[0.036, 0.053]$	$+0.034$
WRN-28-10	D_N	$+0.324$	$[0.314, 0.332]$	$+0.313$
WRN-28-10	D_D	$+0.067$	$[0.057, 0.077]$	$+0.056$

On both backbones the preferred auxiliary reverses between the two error types: all four effects and their Bonferroni lower bounds are positive. The protocol's crossover category nevertheless fails because it requires every effect to reach 0.05, whereas ResNet-18's digit advantage is 0.044. WRN-28-10 meets every criterion. The frozen protocol therefore records a *positive interaction without a crossover*, and Table 1 reports that label. We keep the observed reversal of the preferred auxiliary separate from the protocol's category, which requires the minimum effect.

Uncertainty and replay scope. They hold the ten fitted heads fixed and resample evaluation records, so they describe uncertainty in the evaluation sample, not variability over newly trained backbones. Cluster identifiers do not exist for any of these sources, so resampling is at the image level and inference is restricted accordingly. Matching equalises a specified model-and-validation proxy, not population total variation or Bayes difficulty. The study does not test equality of the full population testing regions. The inference was replayed from the frozen features and fitted heads: confirmation and target artifacts were byte-identical, including the bootstrap draws. The full-grid controls also replay from these inputs. This does not constitute a fresh reproduction of backbone training or head fitting. Archive and checkpoint hashes, stage commands and a replay record state those boundaries explicitly. The large input files are not committed to the repository.

Z.1 Descriptive controls across the complete dilution grid

The protocol specified within-source dilution controls. Table 36 reports all fitted values from the revised grid, including every original value, without selecting a new primary comparison. It gives absolute target AUCs, averaged over the same ten fitted heads per cell, rather than only the selected differences. These are descriptive results on the already-opened evaluation data. No additional significance claim is made.

For a fixed score, dilution only rescales the auxiliary AUC toward $\frac{1}{2}$, by the identity displayed above. At the population optimum, $Q_\lambda/P = (1-\lambda) + \lambda Q/P$ preserves the ranking for positive λ. Neither identity asserts that refitted, regularized linear heads preserve rankings. Across the full grid and fitting replicates, changing dilution within a source changes a target AUC by as much as

0.1688 relative to that source's undiluted head. The minimum Spearman correlation to that head is 0.5270, computed on the combined clean and corresponding target records, with average ranks for ties. The rank summary was specified after the primary outcomes and is descriptive. It measures changes among these fitted scores, not variation over newly trained backbones or an invariant ordering. Source style, semantic affinity and finite-model effects remain entangled. The selected contrasts recomputed here agree with every committed primary point estimate. These descriptive results do not alter the failed primary strong criterion.

	ResNet-18				**WRN-28-10**			
	Natural auxiliary		Letter auxiliary		Natural auxiliary		Letter auxiliary	
λ	CIFAR-100	SVHN	CIFAR-100	SVHN	CIFAR-100	SVHN	CIFAR-100	SVHN
0.25	0.8330	0.7084	0.5332	0.7650	0.8374	0.5785	0.5551	0.7271
0.30	0.8349	0.7116	0.5345	0.7691	0.8398	0.5811	0.5550	0.7274
0.35	0.8367	0.7149	0.5358	0.7731	0.8420	0.5837	0.5548	0.7278
0.40	0.8383	0.7184	0.5370	0.7772	0.8440	0.5864	0.5546	0.7281
0.45	0.8399	0.7220	0.5382	0.7815	0.8460	0.5892	0.5543	0.7285
0.50	0.8413	0.7258	0.5395	0.7861	0.8479	0.5922	0.5539	0.7290
0.55	0.8427	0.7298	0.5408	0.7910	0.8497	0.5955	0.5535	0.7295
0.60	0.8440	0.7341	0.5422	0.7965	0.8515	0.5993	0.5530	0.7302
0.65	0.8452	0.7387	0.5437	0.8026	0.8534	0.6035	0.5524	0.7311
0.70	0.8464	0.7436	0.5455	0.8096	0.8552	0.6085	0.5517	0.7323
0.75	0.8475	0.7489	0.5474	0.8177	0.8572	0.6144	0.5508	0.7339
0.80	0.8486	0.7547	0.5496	0.8274	0.8593	0.6217	0.5495	0.7365
0.85	0.8495	0.7609	0.5524	0.8391	0.8615	0.6310	0.5478	0.7406
0.90	0.8503	0.7677	0.5559	0.8538	0.8639	0.6430	0.5451	0.7480
0.95	0.8509	0.7751	0.5612	0.8717	0.8664	0.6590	0.5395	0.7654
1.00	0.8510	0.7832	0.5121	0.8497	0.8687	0.6807	0.5103	0.8886

Table 36: Complete revised dilution grids for both backbones. Columns give absolute target AUCs for natural and letter auxiliary heads. Each row uses a common dilution for display. The primary comparison uses separately selected mixtures, and no row constitutes a new confirmatory test.

AA SIMULATION DETAILS

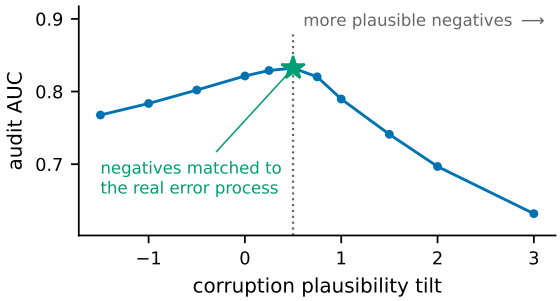

Figure 8: Audit AUC as the corruption plausibility tilt increases, with its signature profile held fixed. AUC peaks where the corruption matches the real error process and falls from 0.832 to 0.632 at the highest-tilt setting, over 400 trials per point. A controlled construction, not a universal claim.

Simulation design. Each claim occupies one of 6×5 cells, indexed by signature $s \in \{0, \dots, 5\}$ (the type pattern of Results 2–4) and plausibility bin $\ell \in \{0, \dots, 4\}$. Let α_{pos} and α_{err} denote correct-claim and error signature profiles. The latter interpolates in log space between the former and an independent Dirichlet draw. Each instance draws the β coefficients uniformly and sets

$$p_{\text{pos}} \propto \alpha_{\text{pos}}(s) e^{\beta_{\text{pos}} \ell}, \quad q_0 \propto e^{\beta_0 \ell}, \quad p_{\text{err}} \propto \alpha_{\text{err}}(s) e^{\beta_{\text{err}} \ell}, \quad q_{\text{hard}} \propto \alpha_{\text{pos}}(s) e^{\beta_{\text{hard}} \ell},$$

with $\beta_{\text{pos}} \in (0.5, 1.2)$, $\beta_0 \in (-1.2, -0.5)$, $\beta_{\text{err}} \in (0.2, 0.8)$ and $\beta_{\text{hard}} \in (0.8, 1.5)$. This models plausible extraction errors: q_0 favors implausible claims, whereas q_{hard} emphasizes plausibility

while retaining the correct-claim signature profile. For the sweep, $\alpha_{\text{pos}} \sim \text{Dirichlet}(2\mathbf{1}_6)$ and independently $\alpha' \sim \text{Dirichlet}(0.4\mathbf{1}_6)$. Interpolation coefficient one sets $\alpha_{\text{err}} = \alpha'$, the endpoint of the specified path. We vary $q \propto \alpha_{\text{err}} \exp(t\ell)$, with p_{err} fixed at $t = 0.5$ on the grid. Increasing t shifts mass toward plausible cells by construction.

The simulation fixes p_{err} by construction and varies the corruption, so every quantity the theory refers to is observable. Over 200 trials with a correctly specified model, worst regret is 0.0 and worst total variation is 1.0×10^{-7}, with no unconverged fits. The budget curve at ridge 10^{-3} recovers a mean 87.1% of the achievable resolution at 25 annotated errors, 94.3% at 50, and 96.6% at 100. The corresponding regret falls from 0.019 to 0.004 over the same range.

Figure 9 measures fidelity and regret bounds on the held-out finite-cell surrogate. Matching improves fidelity, but the bounds remain loose. Result 5b additionally requires both complete scores to be constant within cells, so it applies to this surrogate and does not certify the deployed correction.

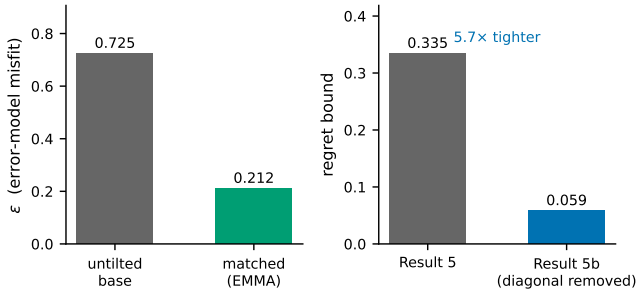

Figure 9: Fidelity and regret bounds on the finite-cell surrogate. Left: matching reduces log-fidelity error ε from 0.725 to 0.212. Right: Result 5 gives 0.335 and Result 5b gives 0.059. Neither bound certifies the deployed correction, whose scores vary within cells.

In 400 simulations driven by the measured error process, 200 annotated errors recover the oracle cell ordering in 99.0% of runs, projecting to 429 annotated claims at the fitting sheet's own observed rate of 140 errors in 300 rows, rather than an observed requirement. That enriched sheet rate, not the calibrated corpus rate of Appendix N, is the denominator here.

Sensitivity of pool-scored regret to contamination. Under Result 7's π-free rule, the fitted tilt and untilted base both attain oracle ordering, with regrets 0.000 and 0.000. Against the contaminated pool, four-cell regret is piecewise constant in π, with the measured contamination near a discontinuity. At $\pi = 0.26$, the untilted base has regret 0.428 and the tilt has zero regret. At $\pi = 0.28$, the tilt's regret is 0.013. Across the tested sweep, the tilt's pool-scored regret never exceeds the base's. The continuous, held-out log-fidelity error ε reported above avoids relying on a single regret value near the discontinuity.

Table 37 is the separate sample-complexity study, simulated at the error process measured on held-out data rather than at a synthetic one. In this simulation, order recovery is essentially complete by 200 annotated errors, while driving ε's 95th percentile below 0.25 needs 400 and improves only slowly after that. These columns describe distinct objectives for annotation planning.

annotated errors	order recovery	ε 95th pct
25	49.5%	1.376
50	75.0%	0.989
100	93.2%	0.558
200	99.0%	0.344
400	99.8%	0.247
800	100.0%	0.184
1,600	100.0%	0.131
3,200	100.0%	0.091

Table 37: Annotation budget and ordering recovery at the measured error process. Order recovery is the fraction of runs whose fitted tilt induces the oracle ordering over cells.