HARM THRESHOLDS AND CERTIFIED MASKS FOR CONSISTENCY REGULARIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Consistency regularization encourages matching predictions for an input and its augmented version. In multi-label and structured prediction, however, an augmentation can preserve some labels while removing the evidence for others. Model confidence measures prediction certainty, whereas selecting outputs for this penalty requires assessing whether augmentation preserves their labels. We derive harm thresholds by distinguishing outputs according to changes in their conditional label probabilities. In a controlled linear model, these thresholds identify, to leading order, the training sample size beyond which a fixed consistency penalty increases prediction risk relative to unregularized training. The crossover arises because the variance benefit decreases with more data while augmentation-induced bias persists. Guided by this analysis, we calibrate output-selection masks on independent validity labels that indicate whether augmentation preserves each output's target. Under i.i.d. calibration sampling and independently constructed validator scores, the certificates bound, at a chosen confidence level, the fraction of selected outputs whose conditional label probabilities change beyond a specified tolerance. Controlled simulations recover the predicted crossover with a median absolute relative deviation of 3.13%. Experiments on document-level relation extraction and named-entity recognition show that penalties applied to every output encourage unchanged predictions after supporting evidence is removed. At comparable selection rates, validated masks keep the mean frequency of this behavior at or below the unregularized level, whereas random masks increase it. These results support selecting where consistency penalties apply according to label preservation at each output.

1 INTRODUCTION

Consistency regularization encourages a model to make matching predictions for an input and its augmented version. In structured prediction, however, different outputs can depend on different evidence. Deleting a sentence from a document may remove the sole support for one relation while leaving other entity pairs unaffected. Applying consistency to every output then rewards an unchanged prediction even for the relation whose evidence has disappeared. The choice of which individual outputs, or *coordinates*, should agree therefore determines which information the model is encouraged to retain.

Existing methods provide several ways to choose consistency targets. Teacher predictions stabilize supervision (Laine & Aila, 2017; Tarvainen & Valpola, 2017), while confidence-based methods select predictions that the model regards as reliable (Sohn et al., 2020; Wang et al., 2023a). For structured outputs, changes in predictions can also indicate which relations an edit affects (Xu et al., 2022). These signals describe prediction reliability or model response. Deciding where to impose consistency additionally requires assessing whether the corresponding prediction target survives the transformation.

This distinction connects the placement of a penalty to its bias–variance tradeoff. Augmentation can reduce estimation variance while introducing bias when the imposed invariance conflicts with the target (Chen et al., 2020; Lin et al., 2024; Yang et al., 2023). Its effects can also differ across outputs (Balestriero et al., 2022). As the training sample grows, a fixed penalty may retain this conflict even as its variance benefit diminishes. Selecting fewer outputs changes both the number of conflicting penalties and the opportunities for useful regularization. We therefore study when a fixed

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.

all-output penalty becomes harmful and how the validity of selected outputs controls the benefit of selective consistency.

We define *coordinate validity* as preservation of an output's conditional label probability within a specified tolerance. This definition supports two distinct analyses. First, under an augmentation that preserves the input marginal distribution, strict prediction invariance incurs a squared-loss excess-risk lower bound determined by the change in conditional label probabilities. Second, in a controlled linear model with a fixed finite penalty, we derive a leading-order sample-size threshold beyond which conflict bias outweighs variance reduction. In a coordinate mixture, we then relate optimized risk to mask *precision*, the fraction of selected outputs that are valid, and *selection rate*, the fraction of all outputs selected. This relation explains why the locations retained by a mask matter alongside their number.

To turn this analysis into a selection rule, we calibrate a score threshold using independent validity labels. Building on finite-sample testing (Bates et al., 2021; Angelopoulos et al., 2021), the calibration controls the proportion of invalid outputs among those selected under the stated sampling conditions. The resulting validated-coordinate consistency (VCC) objective retains supervision on the original input and applies the consistency penalty at the selected coordinates. In the structured-prediction experiments, evidence-based rules select outputs whose evidence footprints remain intact, and held-out evidence annotations assess the precision of those selections.

The experiments test whether selecting valid locations changes how a model responds to removed evidence. On DocRED, all-output consistency increases the fraction of violated coordinates with a prediction change below 0.1 by 19.5 percentage points relative to supervised training (Table 2). A mask that selects coordinates at random at a matched rate retains this excess, whereas the validated mask removes it. The DocRED data-size study further shows that increasing training data does not remove the observed response failure over the evaluated range (Section 6.6). These comparisons connect the empirical effect to penalty placement.

Our contributions are as follows.

- We formalize coordinate validity and distinguish the risk imposed by strict invariance from the sample-size-dependent effect of a finite consistency penalty (Section 3 and Theorem 3).
- We connect the benefit of selective consistency to mask precision and selection rate, and construct calibrated masks for the VCC objective (Corollary 5).
- We evaluate how penalty placement affects predictions after evidence deletion, using matched-rate masks and data-size comparisons to separate selection quality from the number of penalties applied (Sections 6.3 and 6.6).

2 Related Work

Consistency regularization. Consistency objectives match predictions under input or network perturbations (Bachman et al., 2014; Laine & Aila, 2017; Tarvainen & Valpola, 2017; Xie et al., 2020; Miyato et al., 2019; Wu et al., 2021). Selection methods use confidence, adaptive thresholds, class frequencies or learned pseudo-label quality (Sohn et al., 2020; Wang et al., 2023a; Xie et al., 2023; Li et al., 2024b). We instead evaluate augmentation validity at each output. Section 6.4 examines a displacement selector whose score depends on the model it regularizes.

Augmentation and invariance theory. Group averaging (Chen et al., 2020) and kernel models (Dao et al., 2019) explain augmentation through its effects on generalization. Transformation-based analyses characterize its regularization effects (Lin et al., 2024; Yang et al., 2023), and scaling studies relate augmentation to additional training data (Geiping et al., 2023). Class-dependent effects show that regularization and augmentation affect categories differently (Balestriero et al., 2022). Our decomposition isolates regularization on outputs whose targets change. Section 6.6 then examines how the corresponding response behavior varies with training size.

Evidence-sensitive structured prediction. Relational self-distillation matches document views with and without masked entity-pair features (Zhang et al., 2023). Entity-replacement tests (Wang et al., 2023b) and CovEReD (Modarressi et al., 2024) examine or enforce invariance to renaming. SIEF estimates relation-specific sentence importance from prediction changes and penalizes changes after low-importance deletions for consistency training (Xu et al., 2022). Complementary-label learning combines augmentation consistency with annotations specifying excluded categories (Gao

et al., 2026). DocRED evidence annotations (Yao et al., 2019) and CoNLL entity spans (Tjong Kim Sang & De Meulder, 2003) let us assess whether an edit preserves the information supporting an individual relation or word label. Our SIEF-style adaptation and rule mask share the deletion and penalty, with different selection criteria and realized rates.

Learning with noisy labels. Loss correction adjusts training on corrupted annotations using class-conditional noise rates or transition matrices (Natarajan et al., 2013; Patrini et al., 2017). Sample-selection methods instead use model losses to select or relabel training examples (Han et al., 2018; Li et al., 2020; Song et al., 2023). Our selection target concerns changes induced by a known augmentation at individual outputs. Validity labels assess those changes directly and support calibration of the selected subset.

Finite-sample selection and abstention. Selective prediction trades acceptance frequency against prediction error (Chow, 1970; El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017). Risk-controlling prediction sets (Bates et al., 2021) and Learn Then Test (Angelopoulos et al., 2021) calibrate decisions using finite data. Conformal p-values (Jin & Candes, 2023) and selective labeling (Huang et al., 2025) control errors among selected items. The comparison of selection units and guarantees appears in Appendix F. We select outputs of augmented inputs and test their validity errors. The errors we control are changes of the conditional target under augmentation, and they enter the coordinate-mixture comparison as mask precision.

3 Coordinate Validity and Consistency Risk

We define validity, compare selective-penalty risk (Section 4), and calibrate the selection mask (Section 5).

3.1 Coordinate validity and masks

An edit can preserve some targets while changing others, motivating output-level validity. Let $x \sim P_X$ be an input in $\mathcal{X}$ with labels $y \in \{0,1\}^K$. Coordinate k is a single output with conditional label probability $\eta_k(x) = \mathbb{E}[y_k \mid x]$. An augmentation A maps x to Ax and may be fixed or drawn from a specified edit law. We define the target probabilities on the evaluated original and augmented input domain. Values outside the original-data support require additional identifying information (Appendix A.6). The predictor $f_\theta : \mathcal{X} \to [0,1]^K$ estimates these probabilities. Consistency training supplements task supervision with a sum of prediction disagreements across outputs (Laine & Aila, 2017; Xie et al., 2020). The discrepancy D is symmetric Bernoulli KL or squared difference. We call this an *all-output consistency penalty*. Its observed tendency to retain predictions after target-altering edits is *collateral near-invariance*. Strict invariance below denotes an equality constraint on a function.

Definition 1 (Coordinate validity)**.** Fix a tolerance $\varepsilon_v \geq 0$. The *validity indicator* of coordinate k for the pair (x, Ax) is $V_k(x, A) = \mathbf{1}\{|\eta_k(Ax) - \eta_k(x)| \leq \varepsilon_v\}$, where $\mathbf{1}$ equals one when its condition holds and zero otherwise. A coordinate with $V_k = 1$ is *valid* for the augmentation. A coordinate with $V_k = 0$ is *violated*.

All probabilities refer to a specified joint coordinate population of (x, A, k). The theoretical mixture samples output coordinates uniformly, while empirical rates pool the eligible coordinate instances uniformly within each named population. For a single-output result, this law is conditioned on that output. Define violation mass $\pi = \mathbb{P}[V_k = 0]$ and squared violation gap $\gamma^2 = \mathbb{E}[(\eta_k(Ax) - \eta_k(x))^2 \mid V_k = 0]$ when $\pi > 0$. The product measures conflict from violated coordinates and uses target probabilities, independently of model predictions.

A validator supplies a score u_k, from which a binary mask $\hat{V}_k$ selects penalties. Practical selection first restricts eligible outputs using a frozen unregularized reference f_{ref}. We write the resulting pool as $\mathcal{K}(x, A)$, allowing dependence on both views. A common choice includes outputs exceeding a probability cutoff ϵ in either view. CoNLL uses the clean-view pool specified in Appendix E. The cutoff ϵ differs from validity tolerance ε_v.

Within any one chosen population, mask precision p is the valid fraction among selections, recall r is the selected fraction among valid outputs, and selection rate s is the fraction of eligible outputs selected. They satisfy $ps = r(1 - \pi)$ (Appendix B.5). The oracle selects exactly the valid outputs.

Comparisons of (π, p, r, s) use the same population, keeping active-pool training rates separate from gold-pool evaluation rates.

3.2 THE CONFLICT FLOOR

We quantify the error that strict invariance imposes when an augmentation changes the conditional target. For one coordinate write $R(f) = \mathbb{E}[(f(x) - y)^2]$. The excess risk is $R(f) - R(\eta) = \mathbb{E}[(f(x) - \eta(x))^2]$. Let $\mathcal{I} = \{f \in L^2(P_X) : f(x) = f(Ax) \text{ a.s.}\}$ be the closed subspace of invariant predictors with finite second moment. To compare approximation errors on both views, we assume marginal preservation.

Assumption 1 (Marginal preservation)**.** $Ax \overset{d}{=} x$ when $x \sim P_X$.

This preserves the input distribution while permitting paired targets to differ. The resulting conflict floor makes the bias in augmentation analyses (Chen et al., 2020; Lin et al., 2024) explicit.

Lemma 2 (Conflict floor)**.** *Under Assumption 1, every $h \in \mathcal{I}$ satisfies $\mathbb{E}[(\eta - h)^2] \geq \frac{1}{4}\,\mathbb{E}[(\eta(x) - \eta(Ax))^2] \geq \frac{1}{4}\,\pi\gamma^2$. Only the second inequality becomes an equality when valid coordinates are exactly preserved ($\varepsilon_v = 0$).*

This approximation error persists with more data, motivating exclusion of violated outputs from the penalty (proof in Appendix A.1).

4 HARM THRESHOLDS AND SELECTIVE REGULARIZATION

A finite penalty trades variance reduction against conflict bias. We locate its risk crossover in a linear model, then compare mask precision and selection rate under a shared shrinkage law.

4.1 THE HARM THRESHOLD

To analyze a finite penalty, we now use a controlled real-valued regression model, replacing the probability-valued predictor with $f_\theta(x) = \langle \theta, \phi(x) \rangle$ with squared loss and squared prediction disagreement. Write $\Delta = \phi(x) - \phi(Ax)$.

Assumption 2 (Simultaneous diagonalizability)**.** In the linear model $f_\theta(x) = \langle \theta, \phi(x) \rangle$, the matrices $\Sigma = \mathbb{E}[\phi(x)\phi(x)^\top]$ and $\Sigma_{\hat{V}} = \mathbb{E}[\hat{V}\Delta\Delta^\top]$ are simultaneously diagonalizable. Both are population matrices. Σ describes the features and $\Sigma_{\hat{V}}$ describes selected cross-view feature differences.

Let $\hat{f}_\lambda$ minimize empirical squared loss plus $\lambda\Omega_{\hat{V}}$ on n examples. The penalty uses the population matrix $\Sigma_{\hat{V}}$, while task loss uses the training sample. The conditional mean is $\eta(x) = \langle \theta^\star, \phi(x) \rangle$ with homoscedastic noise variance σ^2. The fixed-dimension moment and inverse-moment conditions for the risk expansion are stated in Appendix A.2. The two-block reduction has an invariant block contributing equally to both estimators and a penalized block of dimension d_1. Here d_1 counts feature directions for one output, not output coordinates. The penalized block has feature eigenvalue $\rho > 0$ and penalty eigenvalue $\mu > 0$. Its true parameter is $\theta_1^\star$. Raising penalty strength lowers the variance along selected directions and raises the bias toward invariant predictions. The spectral risk decomposition and its two-block interpretation are given in Appendices A.2 and B.6.

Theorem 3 (Harm threshold)**.** *Under Assumptions 1 and 2, the expansion conditions in Appendix A.2, and the two-block reduction, fix $\lambda > 0$ with $\hat{V} \equiv 1$ and $\pi\gamma^2 > 0$. Let $\Delta_n(\lambda)$ be the excess risk of indiscriminate consistency minus the excess risk of unregularized least squares. Then*

$$\Delta_n(\lambda) = \Big(\frac{\lambda\mu}{\rho + \lambda\mu}\Big)^2 \Big[B_\infty - \frac{\sigma^2 d_1}{n}\Big(1 + \frac{2\rho}{\lambda\mu}\Big)\Big] + O\big(B_\infty n^{-1} + n^{-2}\big),$$
$$B_\infty = \rho\|\theta_1^\star\|^2 \;\geq\; \tfrac{1}{4}\pi\gamma^2, \qquad n^\star(\lambda) := \frac{4\sigma^2 d_1}{\pi\gamma^2}\Big(1 + \frac{2\rho}{\lambda\mu}\Big). \tag{1}$$

When valid coordinates are exactly preserved, $B_\infty = (\rho/\mu)\pi\gamma^2$. In general the displayed lower bound still holds. The bracket is positive for $n > n^\star(\lambda)$, giving a sufficient leading-order harm threshold. The remainder implies $\Delta_n(\lambda) > 0$ for all sufficiently large n.

The model's conflict bias persists at fixed λ, while its variance benefit decreases as $1/n$. Under exact preservation and $\mu = 4\rho$, $n^\star$ equals the leading-order crossover. Otherwise the crossover depends on

B_∞ (Appendix A.3). Section 6.6 tests the related evidence-response behavior in real tasks, whose F1 metric differs from this squared-risk calculation.

4.2 How precision and selection rate affect optimized risk

We compare masks and all-output consistency after optimizing both penalty weights. A homogeneous coordinate mixture holds shrinkage fixed across outputs, separating selection quality from quantity.

Assumption 3 (Coordinate-mixture reduction)**.** Each coordinate has unregularized excess risk $\mathcal{R}_0 = \sigma^2 d_1/n$. A selected coordinate shares the two-block shrinkage factor $w = \rho/(\rho + \lambda\mu)$. A selected valid coordinate is exactly preserved and hence has zero conflict bias and excess risk $\mathcal{R}_0 w^2$. A selected violated coordinate has excess risk $\mathcal{R}_0 w^2 + (1-w)^2 B_\infty^v$, where $B_\infty^v > 0$ is the limiting bias of forcing that coordinate to agree. Unselected coordinates retain risk $\mathcal{R}_0$. The mask selects a fraction $s \in (0, 1]$ with precision p.

This exact-preservation mixture uses $B_\infty^v = \rho\gamma^2/\mu$ and $B_\infty = \pi B_\infty^v$. The unselected, selected-valid and selected-violated groups have weights $1 - s$, sp and $s(1-p)$, respectively. Summing their risks gives

$$\mathcal{E}_{\text{agg}}(w) = (1-s)\mathcal{R}_0 + s\mathcal{R}_0 w^2 + s(1-p)(1-w)^2 B_\infty^v.$$

Let $\mathcal{E}_{\text{agg}}$ denote mean excess risk across coordinates. The mixture derivation and the role of selected feature directions are given in Appendix B.7.

Theorem 4 (Aggregate risk under an imperfect validator)**.** *Under Assumption 3, the minimum mean excess risk after optimizing the penalty strength is*

$$\mathcal{E}^\star_{\text{agg}}(s, p) = \mathcal{R}_0 - \frac{s\mathcal{R}_0^2}{\mathcal{R}_0 + (1-p)B_\infty^v}, \tag{2}$$

The subtracted term is the improvement over unregularized risk $\mathcal{R}_0$.

Within feasible masks, increasing s regularizes more outputs and increasing p reduces bias per selection. Feasibility requires $sp \le 1 - \pi$ and $s(1-p) \le \pi$. The optimal weight and marginal gains are derived in Appendix A.5.

Let $\alpha \in (0, 1)$ be the allowed fraction of violated selections. A precision certificate gives $p \ge 1 - \alpha$.

Corollary 5 (When masking is worth it)**.** *Under the coordinate mixture, a mask certified at $p \ge 1 - \alpha$ has strictly smaller optimized risk than indiscriminate consistency if*

$$s > \frac{\mathcal{R}_0 + \alpha B_\infty^v}{\mathcal{R}_0 + \pi B_\infty^v}. \tag{3}$$

Here the right-hand side is the required selection rate at the certified precision level. It compares the mask with an indiscriminate penalty whose strength has also been optimized.

The condition is nonempty only when $\alpha < \pi$. The certificate bounds actual precision p from below. Random selection has $p = 1 - \pi$ and cannot improve on this optimized comparator for $s \le 1$. The identification conditions for recovering validity from task labels are stated in Appendix A.6. Ordinary task labels describe x alone, so we certify precision with independently verified validity labels.

5 A Finite-Sample Certificate for the Mask

We calibrate validator scores using independent validity labels to bound errors among selected outputs.

5.1 Calibrating the selection threshold

We need a population precision bound for a threshold chosen from finite data, since the observed valid fraction alone does not account for sampling uncertainty or threshold selection. Let $u_k(x, Ax) \in \mathbb{R}$ be a validator score with larger values indicating greater validity. Write u, V for a sampled coordinate. Threshold τ selects $u \ge \tau$ with population precision $\text{prec}(\tau) = \mathbb{P}[V = 1 \mid u \ge \tau]$ when this event has positive probability. Fix a finite grid $\tau_1 > \cdots > \tau_T$ and its order before drawing the calibration sample.

Assumption 4 (Calibration sampling)**.** The calibration pool $\mathcal{C}$ contains m i.i.d. coordinates from the deployment population with verified validity labels, and the validator score is constructed from

data independent of $\mathcal{C}$. The threshold grid and its order are also fixed independently of the entire calibration sample, including its scores.

At threshold τ, let $N_\tau = \sum_{i\in\mathcal{C}} \mathbf{1}\{u_i \geq \tau\}$ and $S_\tau = \sum_{i\in\mathcal{C}} V_i \mathbf{1}\{u_i \geq \tau\}$. Compute the one-sided Clopper–Pearson lower bound $L_\delta^-(\tau)$ from these selected and valid counts (Clopper & Pearson, 1934). Here α bounds the allowed invalid fraction and δ is the certification error level. Set $L_\delta^- = 0$ if $N_\tau = 0$ or $S_\tau = 0$.

Fixed-sequence testing evaluates the thresholds in their fixed descending order (Bates et al., 2021; Angelopoulos et al., 2021). Accept each threshold whose lower bound reaches $1-\alpha$ and stop at the first failure. Return the preceding accepted threshold, or $\perp$ if the first test fails. If all tests pass, return τ_T.

Theorem 6 (Certified mask precision)**.** *Under Assumption 4, the fixed-sequence rule satisfies* $\mathbb{P}[\hat{\tau} \neq \perp,\ \mathrm{prec}(\hat{\tau}) < 1-\alpha] \leq \delta$*, where the probability is over the calibration sample.*

This controls the joint event of returning a mask and missing the precision target (proof in Appendix A.7). In contrast, substituting empirical $s, \pi, \mathcal{R}_0, B_\infty^v$ into Eq. (3) gives a mixture-risk diagnostic. The precision certificate alone does not certify that estimated improvement.

5.2 Validated-coordinate consistency

A returned threshold defines $\hat{V}_k^+ = \mathbf{1}\{u_k \geq \hat{\tau}\}$ on the eligible pool, with zero mask outside it. If calibration returns $\perp$, the procedure returns no certified mask and stops before masked training. Algorithm 1 gives this one-sided procedure.

The rule validator R-VAL checks whether an edit intersects the output's evidence footprint, using mentions or lexical evidence available independently of the trained predictor. Dataset evidence annotations supply separate operational validity labels for evaluation. In DocRED these are annotated relation-support sentences; CoNLL uses gold entity spans. They instantiate evidence preservation, while V in the theory denotes conditional-probability preservation. Appendix B.2 specifies the rule, learned and human scores.

The training objective retains original-input task supervision and adds the validity-masked consistency term

$$\mathcal{L}_{\mathrm{main}} = \mathcal{L}_{\mathrm{task}} + \lambda \Omega_{\hat{V}^+}, \tag{4}$$

where $\Omega_{\hat{V}^+}$ uses $k \in \mathcal{K}(x, A)$ with the task's specified minibatch reduction. Selecting every eligible coordinate recovers the all-output comparator, and $\lambda = 0$ recovers ERM. Section 6.3 reports descriptive gold-pool precision for the real-task binary rules and explains their sampling difference from Theorem 6.

The optional two-sided objective adds supervision transfer and a sensitivity margin (Appendix B.8). Their weights are zero in the main real-task comparisons. A mixture-model plug-in weight and its parameter sources are given in Appendix B.3. Training adds one score evaluation per eligible coordinate, with cost depending on the validator and its features. Reference-model evaluation, feature construction and one-time calibration are accounted for separately in Appendix B.3.

6 Experiments

We first test the risk crossover and certificate error control where validity is known. We then ask whether penalty location changes responses to removed evidence on two real tasks. Displacement detection and preserved-evidence controls examine that interpretation, while data-size scans test its persistence. These behavioral measurements complement the controlled squared-risk analysis.

6.1 Controlled risk and certificate reliability

Risk crossover. Family A varies feature-reflection probability across 15 settings. Family B varies the fraction of affected outputs across four settings. Each uses a fixed penalty $\lambda = 1$ and 60 independent designs per sample size. The empirical crossover interpolates where noise-averaged squared risk minus unregularized risk changes sign. Family A uses the population penalty and its two-block crossover prediction. Family B uses an empirical penalty and compares against $n^\star$ under $\mu = 4\rho$ (Appendix C.2).

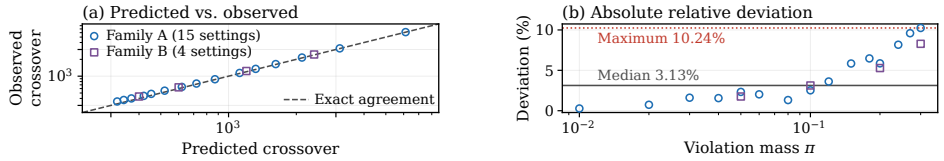


Figure 1: **Controlled risk crossovers.** Left: observed versus predicted sample size, with equality dashed. Right: deviation against violation mass π. The median absolute relative deviation is 3.13%, the maximum 10.24%. Family A uses the two-block prediction and family B $n^\star$.

Table 1: **Error control and abstention are separate outcomes.** Each setting uses $20{,}000$ independent pools of $m = 2{,}000$ coordinates, $T = 50$, $\pi = .2$ and $\alpha = \delta = .05$. Error means returning population precision below .95. Rates and the normal-approximation upper endpoints are percentages. Appendix C.3 specifies grids and interval calculation.

Score/grid	Valid thresholds	Error	Approx. upper	Abstain
Informative	39/50	4.21	4.49	0
Boundary-tight	9/50	2.58	2.80	0
Uninformative	0/50	0.00	0.01	100

Figure 1 places the typical deviation near 3%, with a maximum near 10%. All measured crossings exceed their predictions, consistent with a leading-order approximation that understates the finite-sample crossover. The separate penalty sweep relates this residual to d/n (Table 6).

Certificate reliability. We next test the probability of *returning an underprecise threshold*, separately from abstention. Informative scores rank validity, while a boundary-tight grid uses the same score distribution with more thresholds near the precision target. An uninformative score supplies the negative control. Table 1 shows error rates and their upper bounds below the nominal 5%, consistent with Theorem 6. Complete abstention in the uninformative setting illustrates the cost of having no certifiable threshold. Error control permits occasional incorrect returns.

6.2 Real tasks and comparison design

DocRED (Yao et al., 2019) uses ATLOP with BERT-base (Zhou et al., 2021; Devlin et al., 2019), sentence deletion and entity-pair–relation coordinates. CoNLL-2003 (Tjong Kim Sang & De Meulder, 2003) uses BERT-base-cased, word masking and word–gold-label coordinates. Evidence annotations label whether an edit touches relation support or an entity mention. These are operational validity proxies. Table 7 specifies the pools and budgets.

ERM supplies the unregularized reference. All-output (UDA-style) (Xie et al., 2020), random-mask and rule-mask objectives share the backbone, edits and fixed consistency weight within each domain. Random selection targets the rule's measured rate. DocRED also includes a SIEF-style displacement selector (Xu et al., 2022). These comparisons test location at fixed strength, whereas Section 4 optimizes strength. Initialization and evaluation edits are paired. Best development-F1 selection and response diagnosis use the same development split.

We measure the undesirable behavior of predicting almost unchanged probabilities after evidence removal: CVR $= \mathbb{P}[|f_\theta(x)_k - f_\theta(Ax)_k| < .1 \mid V_k = 0]$. Here V is the evidence proxy and both views use the gold-label probability. Lower CVR means less altered-evidence near-invariance. Relation micro-F1 and entity F1 measure task performance separately.

6.3 Penalty location and selected-output validity

Table 2 compares location while making selection quantity comparable. Random masks exceed rule masks by 19.5 percentage points on DocRED and 33.3 on CoNLL, with positive differences on every paired run. Random-minus-all-output intervals include zero, so these experiments show no stable reduction from random thinning at the tested rates. The optimized mixture model separately explains why random selection cannot improve its comparator when $p = 1 - \pi$.

To check what the rule selects, Table 13 reports valid/invalid selected counts, precision and coordinate-binomial lower bounds for both gold pools. Observed precision exceeds .97 in both domains. These are descriptive summaries: coordinates share documents or sentences, and the development split also

Table 2: **Selection location separates response behavior at comparable penalized fractions.** Means (SD) over five DocRED and three CoNLL runs. s_{tr} is the measured training fraction, distinct from gold-pool selection. Paired intervals resample whole run pairs 20,000 times. CVR is the altered-evidence near-invariance rate at $\kappa = .1$ (lower is better).

	DocRED			CoNLL-2003		
Method	F1↑	CVR↓	s_{tr}	F1↑	CVR↓	s_{tr}
ERM	61.01 (0.25)	.276 (.003)	–	93.93 (0.36)	.487 (.011)	–
All-output	60.66 (0.21)	.470 (.004)	1.000	90.99 (0.07)	.718 (.004)	1.000
SIEF-style	60.84 (0.19)	.289 (.002)	.900	–	–	–
Random mask	60.89 (0.15)	.468 (.005)	.680	90.65 (0.45)	.714 (.009)	.716
VCC (R-VAL)	60.90 (0.27)	.274 (.003)	.696	91.97 (0.30)	.382 (.017)	.716
Paired CVR differences in percentage points: mean [95% bootstrap interval]						
All-output − ERM	+19.5 [19.0, 19.8]			+23.1 [21.6, 24.6]		
Random − rule	+19.5 [19.0, 20.0]			+33.3 [30.6, 35.7]		
Random − all-output	−0.19 [−0.74, 0.36]			−0.34 [−0.80, 0.04]		

Table 3: **Displacement detection and validity selection have opposite positive classes.** Top: displacement detects *violated* evidence on the indicated fitted DocRED model. Bottom: masks select *valid* evidence on the 12,162-coordinate gold pool. Both report means over five runs. The gold-pool mask is a restriction of the all-coordinate evaluation mask, so its selection rate need not equal the training fraction in Table 2.

Model supplying score	Violation precision	Violation recall
ERM	.824	.698
SIEF-style	.299	.973
VCC (R-VAL)	.600	.774
Valid selection on the gold pool: precision p and rate s		
SIEF-style	.956	.160
Rule	.977	.624

selects the model. The independent-calibration guarantee of Theorem 6 applies to the controlled study's sampling design.

6.4 DISPLACEMENT SCORES ON THE MODELS THEY REGULARIZE

The SIEF-style comparison tests a particular model-dependent selector: a stop-gradient batch quantile retains the lowest 90% of training displacements. Table 3 separates two questions. Can displacement detect *violated* evidence, and does the complementary selector retain *valid* evidence? Restricting the evaluation mask to gold-positive coordinates changes its selection rate, explaining why that rate differs from the training fraction.

Violation-detection precision drops from .824 on ERM to .299 on the SIEF-style model. This association is consistent with feedback through the learned score. It does not isolate that feedback as a causal intervention. SIEF-minus-rule CVR is positive on all five runs (Table 9), while the smallest attainable two-sided sign-flip value is .0625. The paired F1 interval spans zero, providing no clear task-score separation.

6.5 RESPONSES TO PRESERVED AND ALTERED EVIDENCE

The placement interpretation also predicts that reduced CVR need not remove agreement on preserved evidence. Figure 2 compares both coordinate groups under the same edit. Preserved-evidence agreement remains high under all four objectives, while the rule lowers near-invariance on altered coordinates. Alongside the random-mask contrast, this supports a role for selection location without identifying every effect of overall regularization strength.

6.6 PERSISTENCE WITH TRAINING SIZE AND TASK-SCORE LIMITS

Figure 3 uses three common seeds at every size. DocRED CVR differences grow across the measured range, while CoNLL means decrease initially and then remain close. Positive excess persists in both

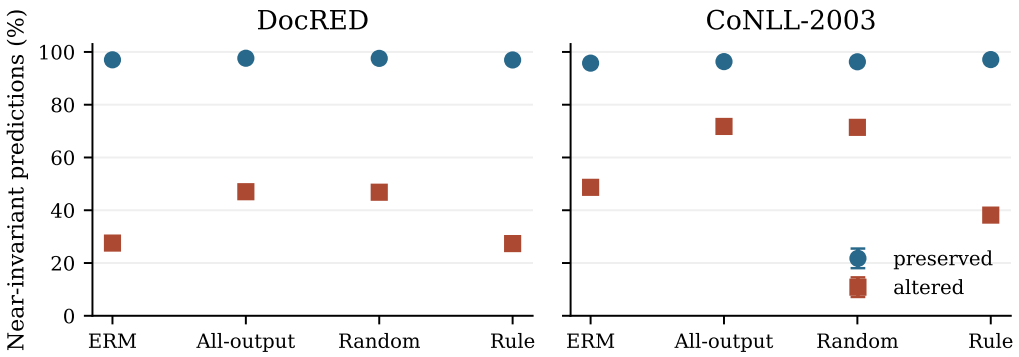


Figure 2: **Agreement remains concentrated on preserved evidence.** Means ± SD over five DocRED and three CoNLL runs at $\kappa = .1$. Preserved/altered pools contain $9{,}019/3{,}143$ and $5{,}979/2{,}236$ coordinates, respectively. The altered rate is CVR. Full tolerances appear in Tables 14 and 15.

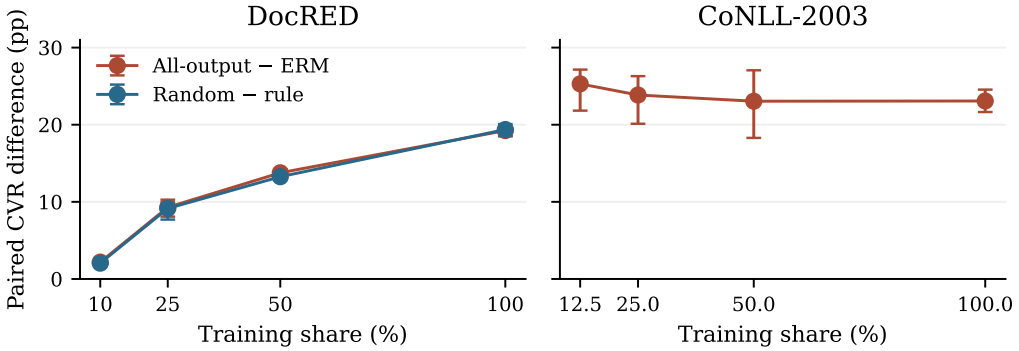


Figure 3: **Persistent CVR excess has different size trends.** Three matched seeds per size, including full size. Bars are 95% percentile intervals from 20,000 whole-pair bootstrap resamples. Task scores and counts are in Tables 12 and 17.

domains. Fixed epochs also increase optimizer updates with dataset size, so these scans combine data and training-budget effects. CVR is not squared risk and cannot locate the theoretical risk crossover.

Task F1 provides a separate limit. DocRED differences are small, and every tested consistency objective trails ERM on CoNLL. The rule has the highest CoNLL F1 among the consistency methods, while ERM remains higher (Table 2). A separate SQuAD 2.0 configuration adds supervision-transfer and null-target sensitivity losses and lowers F1 relative to ERM (Appendix B.4). Thus the real-task evidence supports lower collateral invariance through rule-based placement, with no consistent task-F1 gain.

The metrics answer different questions: CVR conditions on edited, evidence-altered coordinates, whereas task F1 evaluates original inputs. A model can therefore change its response to evidence removal without changing enough original-input decisions to move F1. The placement comparison identifies a behavioral effect under the specified edits and budgets, with task accuracy assessed by its own paired comparisons.

7 Conclusion

Consistency penalties can reward agreement after an edit changes an output's target. Our controlled model characterizes when that conflict bias outlasts the variance benefit, and when mask precision and selection rate improve the optimized risk. Independent calibration controls the probability of returning an underprecise mask. The real-task comparisons then isolate a practical role for penalty location: evidence-based rules reduce altered-evidence near-invariance at comparable selection rates, while task F1 shows no consistent gain. These results make output-level validity a measurable criterion for deciding where consistency regularization should act.

AI USE STATEMENT

In this work, we used generative AI tools for tasks with required disclosure. Generative AI tools helped develop the theoretical models and conceptual frameworks, helped formulate the mathematical claims, provided critical ingredients for proving those claims, and assisted in the writing of the proofs. The mathematical claims, the proofs, and every experimental result were checked and verified by the authors. We additionally used generative AI tools for tasks with recommended disclosure, namely to aid and polish the writing and to edit the paper for readability. We have reviewed all AI-assisted work, including the mathematical derivations and the text. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Verónica Álvarez, Santiago Mazuelas, Steven An, and Sanjoy Dasgupta. Reliable programmatic weak supervision with confidence intervals for label probabilities. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(12):11534–11547, 2025. doi: 10.1109/TPAMI.2025.3597508.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candes, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021.

Anastasios Nikolas Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=33XGfHLtZg.

Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2020. doi: 10.1109/IJCNN48605.2020.9207304.

Martin Arjovsky, Leon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Randall Balestriero, Léon Bottou, and Yann LeCun. The effects of regularization and data augmentation are class dependent. In *Advances in Neural Information Processing Systems 35*, pp. 37878–37891, 2022.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34, 2021.

Emmanuel J. Candès, Andrew Ilyas, and Tijana Zrnic. Probably approximately correct labels. *arXiv preprint arXiv:2506.10908*, 2025. URL https://arxiv.org/abs/2506.10908.

Shuxiao Chen, Edgar Dobriban, and Jane H. Lee. A group-theoretic framework for data augmentation. *Journal of Machine Learning Research*, 21(245):1–71, 2020.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

Tri Dao, Albert Gu, Alexander Ratner, Virginia Smith, Chris De Sa, and Christopher Re. A kernel theory of modern data augmentation. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

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*, pp. 4171–4186, 2019. doi: 10.18653/v1/N19-1423.

Bat-Sheva Einbinder, Liran Ringel, and Yaniv Romano. Semi-supervised risk control via prediction-powered inference. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(12):11080–11090, 2025. doi: 10.1109/TPAMI.2025.3594263.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

Yi Gao, Yuan-Yuan Meng, Miao Xu, and Min-Ling Zhang. Data augmentation with regularization for multi-labeled complementary label learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 48(3): 2368–2379, 2026. doi: 10.1109/TPAMI.2025.3626850.

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

Jonas Geiping, Micah Goldblum, Gowthami Somepalli, Ravid Shwartz-Ziv, Tom Goldstein, and Andrew Gordon Wilson. How much data are augmentations worth? an investigation into scaling laws, invariance, and implicit regularization. In *International Conference on Learning Representations*, 2023.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, 2018. URL https://arxiv.org/abs/1804.06872.

Huipeng Huang, Wenbo Liao, Huajun Xi, Hao Zeng, Mengchen Zhao, and Hongxin Wei. Model-agnostic selective labeling with provable statistical guarantees. *arXiv preprint arXiv:2510.14581*, 2025.

Ying Jin and Emmanuel J. Candes. Selection by prediction with conformal p-values. *Journal of Machine Learning Research*, 24(244):1–41, 2023.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*, 2020.

Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations*, 2017.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 107–117, 2016.

Junnan Li, Richard Socher, and Steven C. H. Hoi. DivideMix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJgExaVtwr.

Juntao Li, Xiaobo Liang, Lijun Wu, Yue Wang, Qi Meng, Tao Qin, Min Zhang, and Tie-Yan Liu. Randomness regularization with simple consistency training for neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(8):5763–5778, 2024a. doi: 10.1109/TPAMI.2024.3370716.

Siyuan Li, Weiyang Jin, Zedong Wang, Fang Wu, Zicheng Liu, Cheng Tan, and Stan Z. Li. SemiReward: A general reward model for semi-supervised learning. In *International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=dnqPvUjyRI.

Julian Lienen, Caglar Demir, and Eyke Hüllermeier. Conformal credal self-supervised learning. In *Proceedings of the Twelfth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 204 of *Proceedings of Machine Learning Research*, pp. 214–233, 2023. URL https://proceedings.mlr.press/v204/lienen23a.html.

Chi-Heng Lin, Chiraag Kaushik, Eva L. Dyer, and Vidya Muthukumar. The good, the bad and the ugly sides of data augmentation: An implicit spectral regularization perspective. *Journal of Machine Learning Research*, 25(91):1–85, 2024.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1979–1993, 2019.

Ali Modarressi, Abdullatif Köksal, and Hinrich Schuetze. Consistent document-level relation extraction via counterfactuals. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 11501–11507, Miami, Florida, USA, November 2024. doi: 10.18653/v1/2024.findings-emnlp.672.

Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789, 2018. doi: 10.18653/v1/P18-2124.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. In *International Conference on Learning Representations*, 2021.

Abhishek Sinha, Kumar Ayush, Jiaming Song, Burak Uzkent, Hongxia Jin, and Stefano Ermon. Negative data augmentation. In *International Conference on Learning Representations*, 2021.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):8135–8153, 2023.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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. URL https://aclanthology.org/W03-0419/.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations: Why and how to pass stress tests. In *Advances in Neural Information Processing Systems 34*, 2021.

Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Yue Fan, Zhen Wu, Jindong Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, Bernt Schiele, and Xing Xie. FreeMatch: Self-adaptive thresholding for semi-supervised learning. In *International Conference on Learning Representations*, 2023a. URL https://arxiv.org/abs/2205.07246.

Yiwei Wang, Bryan Hooi, Fei Wang, Yujun Cai, Yuxuan Liang, Wenxuan Zhou, Jing Tang, Manjuan Duan, and Muhao Chen. How fragile is relation extraction under entity replacements? In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pp. 414–423, Singapore, December 2023b. doi: 10.18653/v1/2023.conll-1.27.

Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. R-Drop: Regularized dropout for neural networks. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Ming-Kun Xie, Jia-Hao Xiao, Hao-Zhe Liu, Gang Niu, Masashi Sugiyama, and Sheng-Jun Huang. Class-distribution-aware pseudo-labeling for semi-supervised multi-label learning. In *Advances in Neural Information Processing Systems 36*, 2023.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Wang Xu, Kehai Chen, Lili Mou, and Tiejun Zhao. Document-level relation extraction with sentences importance estimation and focusing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 2920–2929, Seattle, United States, July 2022. doi: 10.18653/v1/2022.naacl-main.212.

Shuo Yang, Yijun Dong, Rachel Ward, Inderjit S. Dhillon, Sujay Sanghavi, and Qi Lei. Sample efficiency of data augmentation consistency regularization. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *PMLR*, pp. 3825–3853, 2023.

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*, pp. 764–777, 2019. doi: 10.18653/v1/P19-1074.

Christopher Yeh, Nicolas Christianson, Adam Wierman, and Yisong Yue. Conformal risk training: End-to-end optimization of conformal risk control. In *Advances in Neural Information Processing Systems*, 2025.

Sangdoo Yun, Seong Joon Oh, Byeongho Heo, Dongyoon Han, Junsuk Choe, and Sanghyuk Chun. Re-labeling ImageNet: from single to multi-labels, from global to localized labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Liang Zhang, Jinsong Su, Zijun Min, Zhongjian Miao, Qingguo Hu, Biao Fu, Xiaodong Shi, and Yidong Chen. Exploring self-distillation based relational reasoning training for document-level relation extraction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13967–13975, 2023. doi: 10.1609/aaai.v37i11.26635.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 14612–14620, 2021.

A Deferred Statements and Proofs

This appendix provides the proofs and structural results supporting Sections 3–5.

A.1 The conflict floor (Lemma 2)

This section proves Lemma 2 from Section 3.

The error arises from incompatible targets. An invariant predictor must give the same answer to views whose conditional label probabilities can differ. Increasing model accuracy on original inputs alone cannot fit both targets perfectly. Marginal preservation makes the approximation errors on the original and edited views comparable under the same input distribution.

> **Lemma 2 (restated).** Under Assumption 1, every $h \in \mathcal{I}$ satisfies $\mathbb{E}[(\eta - h)^2] \geq \frac{1}{4}\,\mathbb{E}[(\eta(x) - \eta(Ax))^2] \geq \frac{1}{4}\,\pi\gamma^2$. Only the second inequality becomes an equality under exact preservation on valid coordinates ($\varepsilon_v = 0$).

Let $h \in \mathcal{I}$, so $h(x) = h(Ax)$ almost surely. Then

$$\eta(x) - \eta(Ax) = \big(\eta(x) - h(x)\big) - \big(\eta(Ax) - h(Ax)\big). \tag{5}$$

We square both sides and apply $(a-b)^2 \le 2a^2 + 2b^2$ to obtain

$$\begin{aligned}\mathbb{E}\big[(\eta(x) - \eta(Ax))^2\big] \le 2\,\mathbb{E}\big[(\eta(x) - h(x))^2\big] \\ + 2\,\mathbb{E}\big[(\eta(Ax) - h(Ax))^2\big].\end{aligned} \tag{6}$$

By Assumption 1, Ax and x have the same marginal law. The two expectations on the right are therefore equal. This gives $\mathbb{E}[(\eta(x) - \eta(Ax))^2] \le 4\,\mathbb{E}[(\eta - h)^2]$. Finally, $|\eta(x) - \eta(Ax)| \le \varepsilon_v$ whenever $V = 1$, so $\mathbb{E}[(\eta(x) - \eta(Ax))^2] \ge \mathbb{P}[V{=}0]\,\mathbb{E}[(\eta(x) - \eta(Ax))^2 \mid V{=}0] = \pi\gamma^2$, with equality when $\varepsilon_v = 0$. □

A.2 Risk decomposition (Theorem 7)

Section 4 states this result in words. The formal version and its proof are here. Consider $\hat{f}_\lambda = \arg\min_f \hat{R}(f) + \lambda\,\Omega_{\hat{V}}(f)$ in the well-specified linear model with noise variance σ^2 and n training examples.

We use the population penalty matrix in Assumption 2. For the random-design expansion, training examples are i.i.d. with fixed feature dimension, $\mathbb{E}[\epsilon \mid \phi] = 0$ and $\mathbb{E}[\epsilon^2 \mid \phi] = \sigma^2$. The population matrix Σ is positive definite. We assume finite eighth feature moments and uniformly bounded eighth moments of $\|\hat{\Sigma}^{-1}\|$ for the sufficiently large sample sizes in the expansion. These moment and inverse-moment conditions justify taking expectations in the second-order expansion around Σ. The population penalty is held fixed while the task-design matrix is sampled.

Theorem 7 (Risk decomposition)**.** *Under Assumptions 1–2 and the sampling and moment conditions above, with* $M_\lambda = (\Sigma + \lambda\Sigma_{\hat{V}})^{-1}$,

$$\mathbb{E}\big[R(\hat{f}_\lambda)\big] - R(f^\star) = \underbrace{\tfrac{\sigma^2}{n}\,\mathrm{tr}\big[M_\lambda \Sigma M_\lambda \Sigma\big]}_{\mathcal{V}(\lambda)} + \underbrace{\lambda^2\,\theta^{\star\top}\Sigma_{\hat{V}} M_\lambda \Sigma M_\lambda \Sigma_{\hat{V}}\theta^\star}_{\mathcal{B}(\lambda)} + O\big(\mathcal{B}(\lambda)n^{-1} + n^{-2}\big). \tag{7}$$

$\mathcal{V}$ is nonincreasing in λ and $\mathcal{B}$ is nondecreasing. Furthermore, $\lim_{\lambda\to\infty}\mathcal{B}(\lambda) = \min_{h\in\mathcal{I}^{\mathrm{lin}}_{\hat{V}}}\mathbb{E}[(\eta - h)^2]$*, where* $\mathcal{I}^{\mathrm{lin}}_{\hat{V}}$ *collects predictors* within the linear hypothesis class *that are invariant on the pairs the mask selects.*

That limit is a projection inside the linear hypothesis class and not onto the whole space of invariant L^2 functions. The gap is why the harm statement of Theorem 3 is routed through the inequality $\mu \le 4\rho$ rather than an equality. The two-block reduction used by Theorem 3 keeps an invariant block the penalty does not touch and a block of dimension d_1 on which Σ has eigenvalue ρ and $\Sigma_{\hat{V}}$ has eigenvalue $\mu > 0$.

The penalized population solution is $\bar{\theta}_\lambda = (\Sigma + \lambda\Sigma_{\hat{V}})^{-1}\Sigma\theta^\star = M_\lambda\Sigma\theta^\star$, whence

$$\bar{\theta}_\lambda - \theta^\star = \big(M_\lambda\Sigma - I\big)\theta^\star = -\lambda M_\lambda \Sigma_{\hat{V}}\theta^\star. \tag{8}$$

The bias contribution to the excess risk is $(\bar{\theta}_\lambda - \theta^\star)^\top\Sigma(\bar{\theta}_\lambda - \theta^\star) = \lambda^2\theta^{\star\top}\Sigma_{\hat{V}}M_\lambda\Sigma M_\lambda\Sigma_{\hat{V}}\theta^\star$, which is $\mathcal{B}(\lambda)$. The variance contribution of the linear smoother M_λ under homoscedastic noise is $\frac{\sigma^2}{n}\,\mathrm{tr}[M_\lambda\Sigma M_\lambda\Sigma]$, which is $\mathcal{V}(\lambda)$. Expanding $\hat{\Sigma}$ around Σ leaves a remainder of order $\mathcal{B}(\lambda)n^{-1} + n^{-2}$. The first-order design fluctuation has zero expectation. Second-order terms contribute $O(\mathcal{B}(\lambda)n^{-1})$ to the bias expansion and $O(n^{-2})$ to variance, under the stated moment conditions. Under Assumption 2 both matrices diagonalize in a common basis, so each eigen-coordinate contributes $\frac{\sigma^2}{n}\rho_j^2/(\rho_j + \lambda\mu_j)^2$ to $\mathcal{V}$ and $\lambda^2\mu_j^2\rho_j\theta_j^{\star 2}/(\rho_j + \lambda\mu_j)^2$ to $\mathcal{B}$. The former is nonincreasing and the latter nondecreasing in λ term by term. As $\lambda \to \infty$ every coordinate with $\mu_j > 0$ is annihilated, so $\bar{\theta}_\infty$ is the Σ-projection of $\theta^\star$ onto $\ker\Sigma_{\hat{V}}$. Equivalently, it is the projection of η onto $\mathcal{I}^{\mathrm{lin}}_{\hat{V}}$ *within the linear hypothesis class*. The limit is not the projection onto the space of all invariant L^2 functions, which is generally a strictly larger set and would give a strictly smaller limiting bias. □

A.3 The harm threshold (Theorem 3)

This section proves Theorem 3 and explains the tight crossover used in Section 6.1.

Full risk comparison for Theorem 3. Fix $\lambda > 0$ and let $\hat{V} \equiv 1$. Under Assumptions 1 and 2, the expansion conditions of Appendix A.2, and the two-block reduction, indiscriminate invariance has strictly smaller leading-order excess risk than unregularized least squares if and only if

$$B_\infty < \frac{\sigma^2 d_1}{n}\Big(1 + \frac{2\rho}{\lambda\mu}\Big), \qquad B_\infty := \rho\,\|\theta_1^\star\|^2, \tag{9}$$

where B_∞ is the limiting bias in the penalized block. The right-hand side is the largest such bias compatible with a variance benefit at weight λ. At equality the leading-order risk terms coincide. Lemma 2 consequently gives the following sufficient condition for strictly larger leading-order risk than unregularized least squares.

$$\pi\gamma^2 > 4\,\frac{\sigma^2 d_1}{n}\Big(1 + \frac{2\rho}{\lambda\mu}\Big). \tag{10}$$

Under this condition the conflict from altered targets outweighs the variance reduction.

Restrict to the non-invariant block of dimension d_1. Here Σ has eigenvalue ρ and $\Sigma_{\hat{V}}$ has eigenvalue μ. The invariant block contributes identically to both estimators and cancels. From the proof of Theorem 7,

$$\mathcal{E}(\lambda) = \frac{\sigma^2 d_1}{n}\,\frac{\rho^2}{(\rho+\lambda\mu)^2} + \frac{\lambda^2\mu^2\rho}{(\rho+\lambda\mu)^2}\,\|\theta_1^\star\|^2, \tag{11}$$

and $\mathcal{E}(0) = \sigma^2 d_1/n$ is the same expansion at $\lambda = 0$. Both expressions are the leading terms supplied by Theorem 7 and carry a remainder of order $B_\infty n^{-1} + n^{-2}$. Therefore $\mathcal{E}(\lambda) < \mathcal{E}(0)$ if and only if

$$\frac{\sigma^2 d_1}{n}\Big(1 - \frac{\rho^2}{(\rho+\lambda\mu)^2}\Big) > \frac{\lambda^2\mu^2\rho}{(\rho+\lambda\mu)^2}\,\|\theta_1^\star\|^2. \tag{12}$$

Multiplying by $(\rho+\lambda\mu)^2 > 0$ and using $(\rho+\lambda\mu)^2 - \rho^2 = \lambda\mu(2\rho+\lambda\mu)$ gives

$$\frac{\sigma^2 d_1}{n}\,\lambda\mu\,(2\rho+\lambda\mu) > \lambda^2\mu^2\rho\,\|\theta_1^\star\|^2, \tag{13}$$

and dividing by $\lambda\mu > 0$ yields $\rho\|\theta_1^\star\|^2 < \frac{\sigma^2 d_1}{n}(1 + 2\rho/\lambda\mu)$, which is (9) with $B_\infty = \rho\|\theta_1^\star\|^2$. For (10), use the identity of Remark 9. $\pi\gamma^2 = \mathbb{E}[(\eta(x) - \eta(Ax))^2] = \theta^{\star\top}\Sigma_{\hat{V}}\theta^\star = \mu\|\theta_1^\star\|^2$ in the two-block model. Thus $B_\infty = (\rho/\mu)\,\pi\gamma^2$. Since $\mu \le 4\rho$ (Remark 9), we get $B_\infty \ge \frac{1}{4}\pi\gamma^2$. Thus $\pi\gamma^2 > 4\frac{\sigma^2 d_1}{n}(1 + 2\rho/\lambda\mu)$ implies that (9) is reversed strictly, which is (10). All inequalities are strict. At equality the two risks are equal, so equality means non-improvement rather than strict harm. Note that B_∞ is the risk gap to the best invariant predictor *within the linear class*. It can strictly exceed the L^2 floor of Lemma 2, which is why (10) goes through the inequality $\mu \le 4\rho$ rather than through an equality. □

Remark 8 (When the threshold is exact)**.** Equations (9) and (1) describe the leading term of $\Delta_n(\lambda)$ supplied by Theorem 7. When the design is held fixed and its Gram matrix has the two-block form, the fitted coefficients are linear in the noise and both displays are exact at every sample size. Under random design the unregularized risk carries an additional factor of order $d/(n-d-1)$, so the displays locate the crossover to leading order. Random design also changes the penalized bias and variance. Family B of Section 6.1 additionally estimates its penalty matrix from the same design. Its measured crossover combines these finite-sample effects and exceeds (1) in all 20 settings of Table 6, by .58% to 12.78%, and the excess increases with d/n at the crossover along both axes of that table. The residual gap is associated with d/n in this family. These observations compare the leading-order prediction with the fitted estimators, whose finite-sample risks include fluctuations in both the task and penalty matrices.

Remark 9 (Conflict identity and tightness)**.** In the two-block reduction the conflict energy is determined and not merely bounded. At $\hat{V} \equiv 1$ $\pi\gamma^2 = \mathbb{E}[(\eta(x) - \eta(Ax))^2] = \theta^{\star\top}\Sigma_{\hat{V}}\theta^\star = \mu\|\theta_1^\star\|^2$, hence $B_\infty = (\rho/\mu)\,\pi\gamma^2$. And $\mu \le 4\rho$ always, since Assumption 1 gives $v^\top\Sigma_{\hat{V}}v = \mathbb{E}[(g(x) - g(Ax))^2] \le 4\,\mathbb{E}[g^2] = 4\,v^\top\Sigma v$ for any $g = \langle v, \phi\rangle$. The floor $B_\infty = \frac{1}{4}\pi\gamma^2$ is thus attained precisely when $\mu = 4\rho$, when the augmentation reverses the non-invariant feature block.

So $(\rho, \mu, \pi, \gamma^2)$ cannot be chosen independently in a simulation or a plug-in computation. Fixing three of them fixes B_∞, and family B of Section 6.1 is parameterized at the tight point $\mu = 4\rho$ so that (1) carries no slack. Family A has $\mu = 4\pi$ and is therefore compared with (9), which (1) exceeds by the factor $1/\pi$ in that family.

A.4 No scalar factorization (Remark 10)

Section 4.2 cannot treat the masked penalty operator as a scalar multiple of the unmasked one.

Remark 10 (No scalar factorization)**.** The triple (p, r, s) does not determine $\mathbb{E}[\hat{V}\Delta\Delta^\top]$, and conditional independence of $\hat{V}$ and Δ given the true validity V does *not* imply $\mathbb{E}[\hat{V}\Delta\Delta^\top] = s\,\mathbb{E}[\Delta\Delta^\top]$.

Let Δ be supported on two orthogonal unit directions e_1, e_2. With probability 0.2 the coordinate is violated and $\Delta = e_1$, and with probability 0.8 it is valid and $\Delta = e_2$. Then $\mathbb{E}[\Delta\Delta^\top] = \mathrm{diag}(0.2, 0.8)$. Let $\hat{V}$ be drawn conditionally on V with $\mathbb{P}[\hat{V}{=}1 \mid V{=}0] = 0.1$ and $\mathbb{P}[\hat{V}{=}1 \mid V{=}1] = 0.9$. Since Δ is a function of V, this $\hat{V}$ is conditionally independent of Δ given V. The selection rate is $s = 0.2 \cdot 0.1 + 0.8 \cdot 0.9 = 0.74$ and the selection precision is $p = 0.8 \cdot 0.9/0.74 = 0.973$. In the basis (e_1, e_2),

$$\mathbb{E}[\hat{V}\Delta\Delta^\top] = \mathrm{diag}(0.2 \cdot 0.1,\ 0.8 \cdot 0.9) = \mathrm{diag}(0.02, 0.72), \tag{14}$$

whereas $s\,\mathbb{E}[\Delta\Delta^\top] = 0.74 \cdot \mathrm{diag}(0.2, 0.8) = \mathrm{diag}(0.148, 0.592)$. The two differ by 0.128 in the largest entry, so conditional independence given the true validity does not deliver the scalar factorization. □

A.5 Aggregate risk under an imperfect validator (Theorem 4, Corollary 5)

This section proves the optimized-risk result and selection-rate condition of Section 4.2.

Theorem 4 (restated). Under Assumption 3, write $C = (1-p)B_\infty^v$ for the average limiting bias among selected coordinates. Then $\mathcal{E}_{\mathrm{agg}}(w) = \mathcal{R}_0 + s[(C - \mathcal{R}_0) - 2wC + w^2(\mathcal{R}_0 + C)]$ is minimized at $w^\star = C/(\mathcal{R}_0 + C)$. The minimum is

$$\mathcal{E}^\star_{\mathrm{agg}}(s, p) = \mathcal{R}_0 - \frac{s\mathcal{R}_0^2}{\mathcal{R}_0 + (1-p)B_\infty^v},$$

where $\mathcal{E}^\star_{\mathrm{agg}}$ is the excess risk after optimizing the penalty strength. The subtracted term is the improvement over unregularized risk $\mathcal{R}_0$. For $p < 1$, the optimum is attained at $\lambda^\star_{\mathrm{agg}} = \rho\mathcal{R}_0/[\mu(1-p)B_\infty^v]$. This weight scales as $1/n$ and as $1/(1-p)$ with the remaining quantities fixed. For $p = 1$, the infimum is approached as $\lambda \to \infty$. Both marginal risk reductions from increasing s and increasing p are strictly positive. The ratio of the precision gain to the selection-rate gain is $sB_\infty^v/[\mathcal{R}_0 + (1-p)B_\infty^v]$. As $p \to 1$, this ratio tends to $sB_\infty^v/\mathcal{R}_0$ and grows linearly with n. With the two-block relation $B_\infty^v = \rho\gamma^2/\mu$, its order is $\Theta(n\gamma^2/(\sigma^2 d_1))$ for fixed s and ρ/μ. At fixed $p < 1$, both marginal gains decay as $\Theta(n^{-2})$ and their ratio tends to $s/(1-p)$.

Corollary 5 (restated). Under the coordinate mixture, the indiscriminate comparator has $s = 1$ and $p = 1 - \pi$. A mask with operating point (s, p) has strictly smaller minimized risk if and only if $s > [\mathcal{R}_0 + (1-p)B_\infty^v]/[\mathcal{R}_0 + \pi B_\infty^v]$. For a mask certified at $p \geq 1 - \alpha$, a sufficient condition is

$$s > \frac{\mathcal{R}_0 + \alpha B_\infty^v}{\mathcal{R}_0 + \pi B_\infty^v}.$$

Here the right-hand side is the required selection rate at the certified precision level. It compares the mask with an indiscriminate penalty whose strength has also been optimized.

Aggregate risk. The mask leaves a fraction $1 - s$ unselected, selects a valid fraction sp, and selects a violated fraction $s(1 - p)$. Assumption 3 therefore gives

$$\begin{aligned}\mathcal{E}_{\mathrm{agg}}(w) &= (1-s)\mathcal{R}_0 + sp\mathcal{R}_0 w^2 \\ &\quad + s(1-p)\big[\mathcal{R}_0 w^2 + (1-w)^2 B_\infty^v\big] \\ &= \mathcal{R}_0 + s\big[(C - \mathcal{R}_0) - 2wC + w^2(\mathcal{R}_0 + C)\big],\end{aligned} \tag{15}$$

where $C = (1 - p)B_\infty^v$. Its derivative in w is $2s[-C + w(\mathcal{R}_0 + C)]$, and its second derivative is $2s(\mathcal{R}_0 + C) > 0$. Thus $w^\star = C/(\mathcal{R}_0 + C)$ and substitution gives (2).

Optimal weight. Solving $w^\star = \rho/(\rho + \lambda\mu)$ gives $\lambda^\star_{\mathrm{agg}} = \rho\mathcal{R}_0/(\mu C)$ for $C > 0$. When $C = 0$, the minimum occurs at $w = 0$, approached as $\lambda \to \infty$. This proves Theorem 4.

Precision, coverage, and the comparator. Differentiating (2) in s and $1 - p$ gives the two marginal expressions in Theorem 4. Their ratio is $sB_\infty^v/(\mathcal{R}_0 + C)$, which at $p \to 1$ equals $s\,nB_\infty^v/(\sigma^2 d_1) = \Theta(n\gamma^2/(\sigma^2 d_1))$ by $B_\infty^v = \rho\gamma^2/\mu$ and $\mu \le 4\rho$, and at fixed $p < 1$ tends to $s/(1 - p)$ while both marginals decay as $\mathcal{R}_0^2 = \Theta(n^{-2})$.

The indiscriminate comparator has $s = 1$ and $p = 1 - \pi$, so its minimized risk is $\mathcal{R}_0 - \mathcal{R}_0^2/(\mathcal{R}_0 + \pi B_\infty^v)$. Comparing with (2) and cancelling $\mathcal{R}_0^2 > 0$ yields the displayed condition of Corollary 5. If $p \ge 1 - \alpha$, then $(1 - p)B_\infty^v \le \alpha B_\infty^v$, and replacing the numerator by this upper bound yields the sufficient condition (3). Since the right-hand side of (3) is below one if and only if $\alpha < \pi$, that inequality is necessary for the certified sufficient region to be nonempty and is not sufficient on its own. Any s below the displayed threshold violates the displayed condition of Corollary 5 even when $\alpha < \pi$. □

A.6 Identification (Proposition 11)

This section supports the discussion of independent validity labels in Section 4.2 by stating identification conditions.

Proposition 11 (Identification)**.** *(i) If the law of Ax is not absolutely continuous with respect to P_X on a set of positive probability, which implies that Assumption 1 fails, then $\pi\gamma^2$ is not identified from the observed law of (x, y) together with the operator A. Distinct completions of η on the augmented support give distinct values, and when A is injective on that set the attainable values fill an interval whose width is at most the probability that the augmentation leaves* $\mathrm{supp}(P_X)$*. (ii) If Assumption 1 holds and A is a known bijection, then V is identified in the infinite-sample limit, but only through pointwise knowledge of η.*

Part (i) covers edits whose output distribution differs from that of natural inputs. A document with a sentence removed is not distributed like a natural document, so the augmented support is not covered by the training distribution and no amount of training data pins down the counterfactual label. Part (ii) says that even in the idealized case, self-supervised estimation of the mask is circular, a plug-in validator inheriting the estimation error of the problem the regularizer was meant to help with.

(i) Let $S = \mathrm{supp}(Ax) \setminus \mathrm{supp}(P_X)$ have positive probability under the law of Ax. The observed law of (x, y) constrains η only on $\mathrm{supp}(P_X)$. Define η on S freely subject to $\eta \in [0, 1]$. Each choice induces a different value of $\mathbb{E}[(\eta(x) - \eta(Ax))^2] = \pi\gamma^2$, so the quantity is not identified. If in addition A is injective on S, then $\eta \circ A^{-1}$ is well defined there and $\eta|_S = \eta \circ A^{-1}$ removes the contribution of S. Placing $\eta|_S$ at whichever endpoint of $[0, 1]$ lies farther from $\eta \circ A^{-1}$ maximizes that contribution at $\mathbb{E}[\mathbf{1}\{Ax \in S\} \max(\eta(x)^2, (1 - \eta(x))^2)]$, which is at most $\mathbb{P}[Ax \in S]$. Continuity of the induced value in the interpolation parameter gives every intermediate value, so the width of the attainable interval vanishes with the uncovered mass. Without injectivity the attainable set can be a strict subinterval, as when A maps the whole support to a single point. (ii) If Assumption 1 holds and A is a known bijection then $\mathrm{supp}(Ax) = \mathrm{supp}(P_X)$ and η is identified everywhere by the observed law, hence so is V. The construction in (i) is unavailable. The estimation of V then requires pointwise estimation of η at two inputs, whose error is of the same order as the estimation problem the regularizer addresses. □

A.7 Certified mask precision (Theorem 6)

This section proves the precision certificate of Section 5.

Theorem 6 (restated). Consider a score, threshold grid and order fixed independently of the entire calibration sample. Starting at τ_1, test $H_t : \mathrm{prec}(\tau_t) < 1 - \alpha$ in descending threshold order. Reject H_t when $L_\delta^-(\tau_t) \geq 1 - \alpha$ and stop at the first non-rejection. If no hypothesis is rejected, return $\hat{\tau} = \perp$. Otherwise return the last rejected threshold as $\hat{\tau}$. Under Assumption 4, this rule satisfies

$$\mathbb{P}[\hat{\tau} \neq \perp,\ \mathrm{prec}(\hat{\tau}) < 1 - \alpha] \leq \delta,$$

where $\hat{\tau}$ is the returned threshold and $\perp$ denotes abstention. The probability is over the draw of $\mathcal{C}$. The bound controls the joint event of returning a threshold and having its population precision below the target. No multiplicity correction is needed across the fixed sequence.

Condition on the scoring rule, grid and order, all fixed independently of $\mathcal{C}$. At threshold τ_t, condition also on the number selected. The selected validity indicators are then i.i.d. Bernoulli with parameter $\mathrm{prec}(\tau_t)$. The one-sided Clopper–Pearson lower bound therefore gives a level-δ test of $H_t : \mathrm{prec}(\tau_t) < 1 - \alpha$ (Clopper & Pearson, 1934). A threshold selecting no coordinates has lower bound zero and cannot be rejected, so the empty-selection case is handled without special treatment.

Test the hypotheses in the fixed order and stop at the earliest non-rejection. A false rejection can occur only at the earliest true null reached after all earlier nulls were correctly rejected. Its probability is at most δ by the level of the individual test, and no multiplicity correction is required because the sequence is fixed in advance and testing stops at the first non-rejection. If no hypothesis is rejected, the procedure returns $\perp$ and the event in the display is empty. Otherwise returning an invalid last rejected threshold entails a false rejection of that first true null. The joint return-and-error event has probability at most δ, which is the claimed bound. □

B Method Details

B.1 The coordinate-mixture reduction

We use Assumption 3 stated in Section 4.2.

B.2 Validator constructions

The rule evaluates the evidence footprints' intersection with the edited region before the consistency loss is computed, allowing outputs from the same edited input to receive different decisions even when their predicted probabilities are similar. Evidence determines the selection. The trained predictor determines the disagreement that is penalized at those selected locations. Because the selection uses an external description of the edit and its affected evidence, improving prediction agreement during training cannot by itself make an excluded coordinate become valid under the rule. This separation preserves the meaning of mask precision as the model changes.

The rule validator R-Val checks whether an edit intersects an output's evidence footprint of relevant mentions or triggers. The learned validator L-Val combines edit-coordinate features using a labelled split disjoint from $\mathcal{C}$. The human validator H-Val obtains assessments of (x, Ax) from annotators. Each uses the same precision criterion. For the learned validator the features describe how the edit intersects the coordinate, including evidence overlap and the change in reference-model logits as well as the size of the edited region. The reference model remains fixed. This construction makes the score independent of the evolving predictor whose consistency loss it will select, so the validity decision is based on information that is not itself being trained to agree across views.

B.3 The procedure

Algorithm 1 gives one-sided calibration followed by the main masked objective. Its calibration sample satisfies Assumption 4. The real-task binary rules are evaluated separately on the annotated pools described in Appendix D.4.

The returned $\hat{s}$ is an empirical calibration fraction, distinct from population s. The mixture-risk comparison is a separate diagnostic and does not change the algorithm's accept/abstain decision. An

Algorithm 1 One-sided validated-coordinate consistency

Require: training data $\mathcal{D}$, edit law A, fixed reference f_{ref} and eligible-pool rule $\mathcal{K}$, validator u, fixed ordered grid $(\tau_1, \ldots, \tau_T)$, levels α, δ, weight λ, task-specific loss reduction

Draw m i.i.d. calibration coordinates independently of u, f_{ref} and the grid; obtain verified labels V
Set $\hat{\tau} \leftarrow \perp$
for $t = 1, \ldots, T$ **do**
    Compute N_{τ_t}, S_{τ_t} and $L_\delta^-(\tau_t)$
    if $L_\delta^-(\tau_t) < 1 - \alpha$ **then**
        break
    end if
    $\hat{\tau} \leftarrow \tau_t$
end for
if $\hat{\tau} = \perp$ **then**
    return $\perp$ ▷ no certified mask; no masked-training output
end if
Record empirical rate $\hat{s} = N_{\hat{\tau}}/m$ and calibration counts
for each minibatch B **do**
    Draw augmented views and form the fixed-reference pools $\mathcal{K}(x, A)$
    Set $\hat{V}_k^+ = \mathbf{1}\{u_k \geq \hat{\tau}\}$ for eligible coordinates and zero elsewhere
    Take a gradient step on Eq. (4) with the specified reduction
end for
return trained θ, threshold $\hat{\tau}$, levels (α, δ), and $(m, N_{\hat{\tau}}, S_{\hat{\tau}}, L_\delta^-(\hat{\tau}), \hat{s})$

empty selected minibatch contributes zero consistency loss. CoNLL averages symmetric categorical KL over selected active words; its frozen clean-view pool is computed once, as described in Appendix E.

The learned validator. L-VAL fits a gradient-boosted scorer on a labelled split disjoint from the certification pool. Its edit–coordinate features include token overlap with the evidence footprint and attribution mass of the edited region under a fixed reference model f_{ref}. They also include the coordinate's marginal frequency, the cross-view logit shift under f_{ref} and the edit size. Scoring against f_{ref} rather than the model being trained keeps the mask outside the training dynamics. If the scorer is refit during training, each refit is followed by certification on a fresh disjoint split and δ is split evenly across refits so the guarantee holds jointly.

The mixture-model plug-in penalty. Substituting the certified lower precision $1 - \alpha$ into the coordinate-mixture optimum gives

$$\lambda_{\mathrm{plug}} = \frac{\rho \widehat{\mathcal{R}}_0}{\mu \alpha \widehat{B}_\infty^v}.$$

Here $\widehat{\mathcal{R}}_0$ and $\widehat{B}_\infty^v$ are supplied estimates of unregularized excess risk and per-violation limiting bias. In the controlled study, the generating parameters specify these quantities, with $\mathcal{R}_0 = \sigma^2 d_1/n$ and $B_\infty^v = \rho\gamma^2/\mu$. Substituting estimates does not preserve an additional finite-sample risk guarantee. The real-task studies instead fix λ in advance within each domain. Their $\hat{\pi}$ is an annotated-pool fraction, and $\gamma^2 = 1$ is a reporting convention for the conflict floor, not a fitted deep-model risk parameter. Appendix D reports the separate sample-size weight comparison.

Cost. Training computes an augmented forward view and a selected disagreement loss. The added selection work is one validator evaluation per eligible coordinate: $O(|\mathcal{K}|c_u)$ when a score costs c_u. A rule score uses an evidence-set intersection. Learned scores additionally require their reference-model features and tree evaluation, while human scores require annotation. Reference evaluation can be cached when the pool and features permit it. Constructing the labelled pool and testing its threshold grid are one-time costs. Sorting scores and accumulating counts evaluates a fixed grid in $O(m \log m + T)$ operations, apart from binomial-bound evaluation.

Table 4: Reported computational measurements. DocRED step times are rounded to milliseconds; its two consistency methods have equal reported peak memory. CoNLL values are mean whole-run times and share the stated peak memory. These measurements concern the rule validator.

Task	Method	Time	Peak GPU memory
DocRED	ERM	118 ms/step	–
DocRED	All-output consistency	210 ms/step	11.6 GB
DocRED	VCC (R-VAL)	210 ms/step	11.6 GB
CoNLL	ERM	37.9 min/run	7.1 GB
CoNLL	All-output consistency	42.0 min/run	7.1 GB
CoNLL	VCC (R-VAL)	52.6 min/run	7.1 GB

B.4 SCOPE OF THE ASSUMPTIONS AND OF THE CERTIFICATE

The analytical results describe three specified regimes: the strictly invariant function class under marginal preservation, the controlled linear finite-penalty model, and the homogeneous coordinate mixture. The independent calibration guarantee additionally requires i.i.d. sampled coordinates, verified validity labels, and a score and ordered grid fixed independently of the complete calibration sample. It controls the probability of returning a threshold with insufficient population precision.

Real-task validity labels use the datasets' evidence sentences or entity spans. They measure operational evidence preservation and inherit the completeness and errors of those annotations. The within-document or within-sentence sampling dependence and model-selection reuse are documented once in Appendix D.4. Their coordinate-binomial calculations describe these annotated pools. The $\gamma^2 = 1$ conflict-floor convention specifies a scale rather than measuring conditional-probability changes.

The DocRED and CoNLL comparisons test the one-sided placement of the consistency term. Section 6.6 discusses their behavioral and task-accuracy results. The optional supervision-transfer and sensitivity terms in Eq. (18) define a broader objective. A separate SQuAD 2.0 configuration combines supervision transfer with null-target cross-entropy on altered coordinates, rather than the probability-margin term in Eq. (18). At unit weights it gives F1 66.93 versus 69.02 for ERM, and differential abstention response 0.0006 versus 0.3248. This task-specific comparison, including its oracle-label variant, limits the transfer of the one-sided placement results to broader objectives.

B.5 COORDINATE POPULATIONS AND PENALTY DEFINITIONS

The violation mass describes how often a target changes and the squared gap describes the size of those changes, allowing equally frequent augmentations to impose different risks when the affected labels carry different amounts of conflict. A tolerance determines which changes count as violations. Exact validity requires the conditional label probability to remain unchanged, whereas a positive tolerance admits small target changes and assigns the remaining conflict to the valid population. The total squared change is at least the contribution from violated coordinates.

This section supports the validity and mask definitions in Section 3 by detailing the coordinate population and the selection quantities.

Consistency regularization supplements task supervision by penalizing prediction differences between the two views (Laine & Aila, 2017; Xie et al., 2020). The penalty that acts on every output is

$$\Omega(\theta) = \mathbb{E}_{x,A}\Big[\sum\nolimits_{k=1}^{K} D\big(f_\theta(x)_k,\, f_\theta(Ax)_k\big)\Big], \tag{16}$$

where D measures prediction disagreement and the expectation averages over inputs and augmentation randomness. We use symmetric Bernoulli KL or squared difference for D. We call this the all-output consistency penalty. It encourages agreement without checking whether the target probability at each output has changed.

The total change is at least $\pi\gamma^2$, with the remaining contribution bounded by $(1-\pi)\varepsilon_v^2$. Exact preservation on valid coordinates makes this lower bound an equality.

These probabilities refer to a specified coordinate population. Precision and selection rate must use that same population in a risk comparison. Conditioning on active predictions changes the denominator relative to measuring all gold outputs. One way to focus selection on outputs with

a prediction signal is an *active coordinate pool* computed from an unregularized reference model f_{ref}. A probability cutoff ϵ defines the pool $\mathcal{K}(x, A) = \{k : \max(f_{\mathrm{ref}}(x)_k, f_{\mathrm{ref}}(Ax)_k) > \epsilon\}$. The reference model remains fixed during consistency training, so the population used to decide where the penalty can act does not change as the trained predictor becomes more or less invariant. The cutoff removes coordinates assigned very low probabilities in both views. Section 6.2 specifies the populations used for reported rates. Appendix E gives the token-level pool and its cutoff for named-entity recognition.

The all-output penalty also acts on violated coordinates. The near-invariance measured after training is a model response to that signal, distinct from the strictly invariant function class. To select where the penalty acts, we use a *validator* $\hat{V}_k(x, Ax) \in \{0, 1\}$ that estimates coordinate validity. The value $\hat{V}_k = 1$ selects a coordinate and $\hat{V}_k = 0$ excludes it. We measure the reliability and extent of this selection by

$$p = \mathbb{P}[V_k{=}1 \mid \hat{V}_k{=}1], \quad r = \mathbb{P}[\hat{V}_k{=}1 \mid V_k{=}1], \quad s = \mathbb{P}[\hat{V}_k{=}1], \tag{17}$$

where p is *precision*, the fraction of selected coordinates that are valid. The *recall* r is the fraction of valid coordinates selected. The *selection rate* s is the fraction selected from the same named coordinate population used for p, r and π. These quantities satisfy $ps = r(1 - \pi)$ whenever the conditional probabilities are defined. Precision controls the frequency of incorrect selections, whereas selection rate measures how many coordinates receive regularization.

The all-output mask uses $\hat{V} \equiv 1$ and therefore has $s = r = 1$. Its precision is $p = 1 - \pi$. The oracle uses $\hat{V} = V$ and achieves $p = r = 1$ with $s = 1 - \pi$. A general validator modifies the original penalty only by selecting outputs.

$$\Omega_{\hat{V}}(\theta) = \mathbb{E}_{x,A}\Big[\sum_k \hat{V}_k(x, Ax) D\big(f_\theta(x)_k, f_\theta(Ax)_k\big)\Big],$$

where $\Omega_{\hat{V}}$ is the population sum restricted by the mask. On a practical eligible pool, the mask is zero outside $\mathcal{K}(x, A)$. The theoretical full-output analysis uses all coordinates. Real-task loss reductions and eligible pools are fixed by the corresponding task protocol. Setting $\hat{V} \equiv 1$ recovers Eq. (16). The next step is to determine the risk caused by retaining violated coordinates in this penalty.

B.6 Linear risk model and threshold interpretation

More training data reduces the uncertainty that consistency can remove, while the penalty continues to enforce agreement on targets that differ under augmentation. This is the source of the crossover at fixed penalty strength. The threshold separates compensation for limited training data from the regime where persistent target conflict costs more than the remaining variance benefit.

The full spectral decomposition and its assumptions are given in Appendix A. In the two-block reduction the invariant directions contribute equally to both estimators. Removing the penalized block entirely incurs limiting bias $B_\infty = \rho\|\theta_1^\star\|^2$.

Equation (1) gives a sufficient threshold for harm in the leading-order risk expression. The factor $\sigma^2 d_1$ measures the estimation variance available for reduction. Larger conflict $\pi\gamma^2$ lowers the threshold. Increasing λ also lowers it through the final factor.

Under exact preservation on valid coordinates, the two-block model satisfies $B_\infty = (\rho/\mu)\pi\gamma^2$ with $\mu \le 4\rho$. Remark 9 derives this relation. At $\mu = 4\rho$, the floor is tight and $n^\star$ equals the leading-order risk crossover in this reduction. Family B of Section 6.1 uses that case. Penalty strength changes shrinkage in every selected direction. Masking instead changes which coordinates incur that shrinkage, motivating the comparison below.

B.7 Mixture risk and marginal improvements

This section supports Theorem 4 and Corollary 5 by detailing the coordinate mixture and the interpretation of its risk terms.

Equation (9) compares an all-output penalty with unregularized training. We now ask whether a mask can improve on that penalty even after both methods optimize their penalty weights. The

comparison must account for both incorrect selections and discarded valid coordinates. We use an explicit coordinate mixture to connect the precision and selection rate in Eq. (17) to these two effects.

In this mixture each coordinate has unregularized excess risk $\mathcal{R}_0 = \sigma^2 d_1/n$. We use $\mathcal{R}_0$ for the scalar risk in Assumption 3, reserving A for the augmentation operator. A selected coordinate shares the shrinkage factor $w = \rho/(\rho + \lambda\mu)$. The value $w = 1$ corresponds to no penalty and $w \to 0$ to an increasingly strong penalty. A selected valid coordinate has risk $\mathcal{R}_0 w^2$. A selected violated coordinate has risk $\mathcal{R}_0 w^2 + (1-w)^2 B_\infty^v$, where $B_\infty^v > 0$ is its limiting bias. An unselected coordinate retains risk $\mathcal{R}_0$. The mask selects a fraction $s \in (0, 1]$ with precision p. This model makes selection affect the number and validity of penalized coordinates while holding their shrinkage law fixed.

Assumption 3 states these requirements in full. They specify the risk calculation for the controlled family of Section 6.1. For a general feature model, the selected feature directions also matter. The construction in Remark 10 illustrates this distinction at selection rate 0.74 and precision 0.973. It compares the actual selected penalty matrix with the unmasked matrix multiplied by the selection rate. The matrices differ even at this high precision, which motivates specifying the coordinate mixture explicitly.

We calculate aggregate risk by weighting the unselected and selected coordinate groups.

$$\mathcal{E}_{\mathrm{agg}}(w) = (1-s)\mathcal{R}_0 + sp\mathcal{R}_0 w^2 + s(1-p)\big[\mathcal{R}_0 w^2 + (1-w)^2 B_\infty^v\big],$$

where $\mathcal{E}_{\mathrm{agg}}$ is the mean excess risk across coordinates. The first term retains the risk of outputs receiving no penalty. The second is the variance remaining on selected valid coordinates. The final term adds variance and bias on selected violated coordinates. Theorem 4 optimizes their sum over the common shrinkage factor.

Appendix A differentiates the quadratic risk to obtain the optimum and the marginal gains. Equation (2) separates the two roles of a validator. Increasing s expands the set of regularized coordinates. Increasing p reduces the bias per selected coordinate and permits a stronger penalty. Near perfect precision the relative value of improving precision increases with sample size. At perfect precision the limiting risk is $(1-s)\mathcal{R}_0$, contributed entirely by unselected coordinates.

Equation (2) provides the optimized risk of both a mask and the indiscriminate comparator. We use it to determine how much selection is needed to justify a mask. Let α be an allowed fraction of violated coordinates among selections, so a precision certificate gives $p \geq 1 - \alpha$.

Appendix A proves the result by comparing the two instances of Eq. (2). Choosing $\alpha < \pi$ makes the required selection rate smaller than one. The mask improves on indiscriminate consistency when its selection rate also exceeds that requirement. Thus a validator should retain enough outputs while reducing the fraction whose targets change. Higher precision means fewer selected outputs pull predictions toward targets that conflict across the edited views. The optimized expression evaluates precision and selection rate under the same risk objective. Their joint assessment rewards reliable selection while retaining the variance benefit available from valid coordinates. Precision alone does not specify the retained benefit.

Applying this rule requires an independently supported precision bound. Ordinary task labels describe x, whereas coordinate validity compares the targets at x and Ax. We therefore obtain verified validity labels on a separate calibration sample. Proposition 11 in Appendix A discusses identification from task labels. The certificate controls precision. The selection-rate condition uses the mixture's remaining parameters to assess its predicted benefit separately.

B.8 Optional two-sided objective

This extension adds supervision transfer and a sensitivity margin to the one-sided objective in Eq. (4).

A validity certificate identifies outputs on which to encourage agreement. A complementary certificate identifies outputs on which the target changes. We denote the thresholds for these two events by $\hat{\tau}^+$ and $\hat{\tau}^-$, respectively. The positive mask is $\hat{V}_k^+ = \mathbf{1}\{u_k \geq \hat{\tau}^+\}$ and certifies validity at level $1 - \alpha$. The negative mask is $\hat{V}_k^- = \mathbf{1}\{u_k \leq \hat{\tau}^-\}$ and certifies violation at the same level. We obtain the latter by applying Theorem 6 to the complementary label $1 - V$ with reversed score $-u$. Each application has its own error probability δ. Using δ on both sides therefore controls the union

of incorrect certificates by 2δ, not δ. A desired joint budget can instead be specified as $\delta_+ + \delta_-$. An abstaining side contributes no selected coordinates. This two-sided extension is separate from Algorithm 1.

We retain the original task loss to learn from the observed labels. The positive mask restricts the invariance penalty to outputs certified for label preservation. It also selects where the original label can be reused as supervision on the augmented input. Exact preservation supports this transfer directly; a positive validity tolerance permits a corresponding target-probability discrepancy. The negative mask selects where a margin term encourages sensitivity to the augmentation. Combining these optional components gives the extended objective

$$\begin{aligned}\mathcal{L} = \mathcal{L}_{\text{task}}\big(f_\theta(x), y\big) \\ + \lambda \sum_{k \in \mathcal{K}(x,A)} \hat{V}_k^+ D\big(f_\theta(x)_k, f_\theta(Ax)_k\big) \\ + \mu_{\text{s}} \sum_{k \in \mathcal{K}(x,A)} \hat{V}_k^+ \ell\big(f_\theta(Ax)_k, y_k\big) \\ + \nu \sum_{k \in \mathcal{K}(x,A)} \hat{V}_k^- \big[m_0 - |f_\theta(x)_k - f_\theta(Ax)_k|\big]_+,\end{aligned} \tag{18}$$

where $\mathcal{L}_{\text{task}}$ is the original-input task loss and ℓ is the coordinate-level supervised loss. The weight λ controls *validated invariance*. The weight $\mu_{\text{s}} \geq 0$ controls *validated-coordinate supervision* on the augmented input. The weight $\nu \geq 0$ controls *sensitivity on violation-selected coordinates*. Certification concerns violation-selection precision. The margin encourages a prediction difference, with neither its magnitude nor direction supplied by that certificate. Its value $m_0 > 0$ specifies the desired magnitude and $[z]_+ = \max(z, 0)$ keeps only a positive margin shortfall. The final term therefore decreases as selected predictions become separated by the required margin. Coordinates selected by neither mask receive only the original task supervision.

Setting $\mu_{\text{s}} = \nu = 0$ gives the masked consistency objective used to isolate penalty placement in Section 6.2. With these weights fixed at zero, setting $\hat{V}^+ \equiv 1$ recovers indiscriminate consistency and setting $\lambda = 0$ recovers unregularized training. Appendices D and E specify the consistency weights for the real-domain comparisons. Algorithm 1 gives the main one-sided procedure. The plug-in choice is in Appendix B.3, and Appendix B.3 separates validator, reference-feature and calibration costs. The additional weights μ_{s}, ν and margin m_0 are inputs only to this extended objective.

B.9 Penalty strength under changing sample size

Equation (1) makes the harm boundary depend on the number of training examples at a fixed penalty strength. A penalty evaluated at one training-set size can cross that boundary as the available variance benefit decreases. The limiting cases and penalty-weight scaling are derived in Appendix A.5.

C Controlled Study

This appendix supports Section 6.1 with the linear comparison and the calibration settings.

C.1 The harm region

Figure 4 draws the boundary of Theorem 3 analytically. No measurement enters it.

C.2 Additional controlled comparisons

Linear families and designs. Both families use independent standard Gaussian feature vectors and noise variance $\sigma^2 = 1$ with penalty strength $\lambda = 1$ and feature eigenvalue $\rho = 1$. Family A has $d = 60$ features and a reflected block of dimension $d_1 = 15$. An augmentation reflects that block when the final feature falls below the Gaussian quantile corresponding to π. The true coefficient vector has squared norm .125 on the reflected block and squared norm 1 on the invariant block. Consequently the squared violation gap is $\gamma^2 = .5$ and the population penalty eigenvalue is $\mu = 4\pi$.

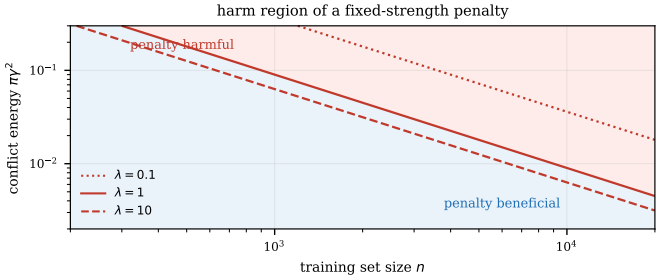


Figure 4: The leading-order harm region of Theorem 3, drawn from the closed form (10). The parameters are σ^2=1 and d_1=15 with ρ=1 and μ=4ρ. Each curve is the boundary at which a fixed-strength penalty stops paying, and the shading corresponds to λ=1. Raising λ moves the boundary down and to the left, so a stronger penalty is harmful over a strictly larger region of $(n, \pi\gamma^2)$.

We evaluate the population penalty matrix for the crossover comparison. The violation probabilities are $\pi \in \{.01, .02, .03, .04, .05, .06, .08, .10, .12, .15, .18, .20, .24, .27, .30\}$. The sample-size grid contains 25 logarithmically spaced values from 120 to 20,000 rounded to distinct integers.

Family B has $d = 40$ features and 200 output coordinates with a shared reflection of a feature block of dimension $d_1 = 10$. It uses $\mu = 4$ and violation fractions $\pi \in \{.05, .10, .20, .30\}$. For each design the violated-coordinate coefficient directions are independent Gaussian vectors normalized to squared norm .125 within the reflected block. The population bias is averaged across the output coordinates. The penalty matrix uses the empirical second moment of the reflected features. Its sample-size grid contains 22 logarithmically spaced values from 80 to 6,000 rounded to distinct integers.

Table 5: Per-setting design counts and sample-size grids for the controlled risk comparison.

Family	Settings	Sample sizes	Range of n	Designs per size
A	15	25	120–20,000	60
B	4	22	80–6,000	60

At every sample size we draw 60 independent design matrices. We compute the noise-averaged squared risk conditional on each design and average its difference from unregularized least squares across designs. The empirical crossover interpolates the sign change on the logarithmic sample-size axis. The predicted crossover uses the exact two-block risk comparison of Appendix A.3 with the family-specific limiting bias. Equation (1) equals that comparison for family B.

Penalty strength and the crossover. Equation (1) makes the crossover proportional to $1 + 2\rho/\lambda\mu$, so a stronger penalty should move it to smaller sample sizes. We repeat the family B comparison at $\lambda \in \{.25, .5, 1, 2, 4\}$ over the same sample-size grid with 60 designs per size. Table 6 reports the relative deviation of the measured crossover from (1) together with the ratio d/n there, and its $\lambda = 1$ column reproduces the family B entries of Figure 1. The measured crossover falls with λ at every violation fraction, and its movement relative to $\lambda = 1$ departs from the predicted ratio by at most 4.81%. Both the deviation and d/n increase along the rows and down the columns of the table. The two ways this family moves d/n at the crossover therefore agree, which separates the residual of (1) from the violation fraction alone. At $\lambda = 1$ the median absolute deviation over the four violation fractions is 4.20%, with a design-bootstrap interval of 3.92% to 4.50% over 5,000 resamples of the designs.

Table 6: Family B under penalty strength. Each cell gives the relative deviation of the measured crossover from (1) in percent, with the ratio d/n at that crossover in parentheses. Every deviation is positive and both quantities increase along the rows and down the columns.

λ	π=.05	π=.10	π=.20	π=.30
.25	.58 (.008)	1.34 (.016)	2.27 (.033)	3.08 (.049)
.5	.72 (.012)	2.05 (.024)	3.75 (.048)	5.69 (.071)
1	1.76 (.016)	3.13 (.032)	5.28 (.063)	8.28 (.092)
2	2.10 (.020)	4.19 (.038)	7.64 (.074)	11.45 (.108)
4	2.42 (.022)	4.29 (.043)	8.90 (.082)	12.78 (.118)

C.3 SCORE DISTRIBUTIONS AND THRESHOLD GRIDS

Each calibration coordinate independently draws V with $\mathbb{P}[V = 0] = \pi = .2$ and then draws its score conditional on validity. In the informative setting $u \mid V = 1 \sim \mathrm{Beta}(5, 2)$ and $u \mid V = 0 \sim \mathrm{Beta}(2, 5)$. Its 50 thresholds are evenly spaced in descending order from .90 to .30. The boundary-tight setting uses these same conditional score distributions. Let $\tau^\star$ satisfy $\mathrm{prec}(\tau^\star) = 1 - \alpha$. Its grid concatenates 8 equally spaced thresholds from .90 to $\tau^\star + .02$ with $\tau^\star$ and then equally spaced thresholds from $\tau^\star - .004$ to .30 for the remaining $T - 9$ entries. In the uninformative setting both conditional distributions are $\mathrm{Beta}(3, 3)$ and the grid is the informative setting's grid. The score is therefore independent of validity in this setting. Because the informative and boundary-tight settings use the same score distributions and differ only in their threshold grids, their comparison isolates how testing near the target precision affects threshold selection when the information available to rank outputs has not changed.

Population precision is evaluated from the conditional Beta survival functions S_v and S_w as $\mathrm{prec}(\tau) = (1 - \pi)S_v(\tau)/[(1 - \pi)S_v(\tau) + \pi S_w(\tau)]$. We apply the fixed-sequence rule separately to each independent calibration pool and compare its selected threshold with this population precision. Drawing validity before the score makes each calibration coordinate a sample from a known joint distribution whose population precision can be evaluated at every candidate threshold, allowing us to compare the reliability of the returned mask with the target even after the threshold has been chosen from that calibration pool. Conditional mean precision is .961 in the informative setting and .967 in the boundary-tight setting. Section 6.1 reports the crossover measurements and incorrect-certificate rates. Table 1 reports the complete threshold counts and abstention probabilities in the main text. The recorded upper endpoints equal $\hat{q} + 1.96\sqrt{\max(\hat{q}, 1/M)/M}$, with $M = 20{,}000$ pools and observed error rate $\hat{q}$. This normal-approximation summary uses a variance upper approximation and a one-event floor at zero. It describes Monte Carlo uncertainty across pools, rather than the precision of an individual returned mask or an exact binomial confidence guarantee.

D DOCRED FULL TABLES

D.1 ALL SIX METHODS

Table 7: Real-task populations and training budgets. Validity labels describe annotated evidence preservation. s_{tr} in Table 2 measures selections from the training penalty pool, while gold-pool selection measures retention of annotated positive coordinates.

	DocRED	CoNLL-2003
Training examples	3,053 documents	14,041 sentences
Prediction coordinate	Entity pair and relation (96 relations)	Word position and BIO label (9 labels)
Edit	Delete one uniformly sampled sentence	Independently mask words with probability .15
Evidence-proxy violation	Deleted relation-support sentence or all mentions of either entity removed	Any word in the gold entity mention masked
Gold pool	12,162 positive relation coordinates	8,215 entity-word coordinates
CVR denominator	3,143 altered coordinates	2,236 altered coordinates
Penalty weight/budget	.1, 30 epochs	1, 2 epochs, batch 48
Evaluation	Development, best relation-F1 epoch	Development, best entity-F1 epoch

Table 8 extends the DocRED block of Table 2 with the evidence-derived oracle. The self-scored SIEF-style selector (Xu et al., 2022) selects coordinates by the model's cross-view logit displacement. It uses batch quantile q=.9 with stop-gradient. The evidence-derived oracle mask is a diagnostic reference rather than a deployable method. All methods use the same consistency weight λ=0.1, thirty epochs and 22,890 optimizer steps. Random selection targets the rule mask's rate, with realized training fractions .680 and .696, respectively. Every configuration at full training size is run with 5 seeds. Contrasts between methods are seed-paired. We report paired means with 95% bootstrap intervals over 20,000 resamples and the number of seeds on which the sign agrees.

The paired validated-minus-random F1 difference is $+0.011$ with interval $[-0.219, +0.187]$ and 3 of 5 seeds positive. The oracle-minus-random difference is $+0.154$ with interval $[+0.021, +0.308]$ and 4 of 5 seeds positive.

Table 8: DocRED at full training size. Task accuracy and collateral violation rate (κ=.1), means over 5 seeds with standard deviations. Task accuracy spans a 0.38 F1 band across all six methods. The collateral column is where they separate. The oracle mask uses evidence-derived validity and is not a deployable configuration.

Method	Ign F1	F1	CVR
ERM (backbone)	59.10 (.26)	61.01 (.25)	.276 (.003)
Indiscriminate (UDA)	58.89 (.17)	60.66 (.21)	.470 (.004)
Self-scored selector (SIEF style)	58.98 (.18)	60.84 (.19)	.289 (.002)
Random mask	59.13 (.19)	60.89 (.15)	.468 (.005)
VCC (R-VAL)	59.12 (.27)	60.90 (.27)	.274 (.003)
Oracle mask	*59.26* (.12)	*61.04* (.09)	*.277* (.007)

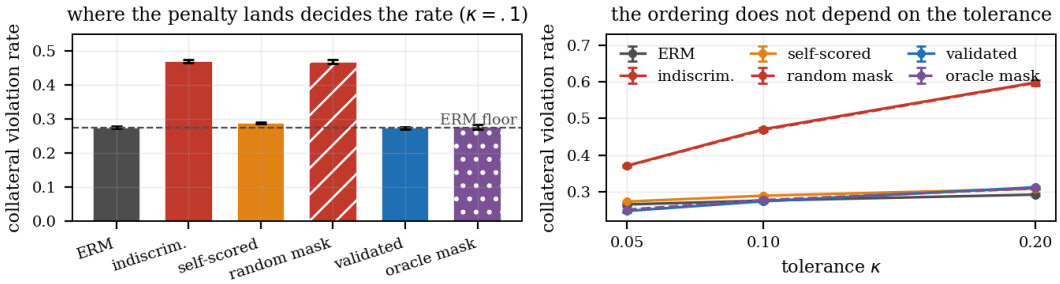


Figure 5: DocRED at full training size, 5 seeds. The left panel shows collateral violation rate at κ=.1, mean $\pm$ standard deviation, with the ERM level marked. The indiscriminate and random-mask methods sit far above it. The validated and oracle methods lie near the ERM reference, while the self-scored selector has a higher collateral rate. Their measured penalised fractions bracket one another. The fractions are .680 for random selection and .696 for the validated mask. They are .783 for the oracle and .900 for the self-scored mask. The indiscriminate method penalises everything. The right panel shows the same quantity at three tolerances κ. Levels move but the ordering does not.

D.2 THE SELF-SCORED SELECTOR AS A DETECTOR

This section supports the precision comparison in Section 6.3 with displacement-score detection and validator operating points.

The SIEF-style training selector uses the lowest $q = .9$ fraction of cross-view displacements within its batch, with stop-gradient through the selection. Its gold-pool diagnosis restricts the recorded all-coordinate evaluation mask to positive gold coordinates. In the saved full-size runs the all-coordinate mask excludes approximately 10% of outputs, while this restriction excludes about 84% of gold coordinates on the SIEF-style model. Hence the .900 training fraction and .160 gold-pool fraction have different denominators. The saved diagnostic records retain assignments and counts, rather than numeric displacement cutoffs. We report that operating point without treating it as a calibrated precision threshold. Table 3 puts violation detection and complementary validity selection in separate blocks.

Resampling both methods jointly preserves their shared initialization within each contrast. The resulting interval measures variation across the fitted models rather than treating the many outputs of a document as independent training replicates.

Table 9: DocRED seed-paired SIEF-style differences. CVR differences are percentage points and F1 differences are F1 points. Intervals use 20,000 whole-pair bootstrap resamples at 95% confidence. Positive counts refer to five seed differences. The final column is the exact two-sided sign-flip value.

Contrast	Metric	Mean	Interval	Positive	p
SIEF − VCC	CVR	1.53	[1.26, 1.79]	5/5	0.0625
SIEF − VCC	F1	-.06	[−.28, .21]	2/5	0.6875
SIEF − ERM	CVR	1.34	[1.01, 1.62]	5/5	0.0625
SIEF − ERM	F1	-.17	[−.50, .16]	2/5	0.4375

Table 10: SIEF-style paired differences for each initialization. CVR differences are percentage points and F1 differences are F1 points. These five pairs are the resampling units of Table 9.

Seed	CVR: SIEF−rule	CVR: SIEF−ERM	F1: SIEF−rule	F1: SIEF−ERM
66	+1.18	+1.46	+0.46	+0.24
67	+1.88	+1.81	-0.32	-0.65
68	+1.18	+0.76	-0.30	-0.55
69	+1.81	+1.46	-0.16	+0.19
70	+1.59	+1.18	+0.01	-0.08

The SIEF-minus-VCC collateral difference is positive on all five matched seeds. Its exact two-sided sign-flip value is .0625. Both paired F1 intervals cross zero and establish neither superiority nor equivalence.

Table 11: The upper block reports the quality of the self-scored displacement score as a detector of evidence-proxy violations on the DocRED gold pool, as a function of which trained model computes it (means over 5 seeds). Its precision approaches the violation base rate ($\hat{\pi} = .258$) precisely on the three methods trained with broad consistency, and the three collapsed rows differ by less than .002 F1. The lower block reports validator operating points in the coordinates of (17), where $\hat{V}{=}1$ means judged valid. The identity $ps = r(1 - \hat{\pi})$ holds to rounding in every row. The indiscriminate row is the degenerate validator $\hat{V} \equiv 1$, with precision $1 - \hat{\pi} = .742$ by definition. The random row carries the same precision because selecting each coordinate independently at a fixed rate preserves the valid fraction of the pool in expectation.

Displacement score computed on	P	R	F1
ERM	.824	.698	.756
VCC (R-VAL)	.600	.774	.676
Oracle mask	.595	.773	.672
Indiscriminate (UDA)	.302	.966	.460
Random mask	.301	.966	.459
Self-scored (its own method)	.299	.973	.458
Validator operating point	$\hat{p}$	$\hat{r}$	$\hat{s}$
Indiscriminate	.742	1.000	1.000
Random (s=.68)	.742	.680	.680
Self-scored (converged)	.956	.207	.160
R-VAL	.977	.822	.624
Oracle	1.000	1.000	.742

D.3 THE DATA-SIZE SCAN

Table 12 supports the DocRED training-size comparison in Section 6.6 with the full paired results and task-score means.

Task F1 measures a different outcome from CVR. Indiscriminate consistency trails ERM by -0.24 and -0.24 F1 at the smaller sizes. Its differences are -0.41 F1 and -0.349 at the larger and full sizes. The rule mask has no consistent mean F1 benefit across sizes. Its differences are -0.64 and $+0.29$ F1 at the smaller sizes with -0.03 and -0.11 at the larger and full sizes. Even the evidence-derived oracle mask sits at -0.35 F1 at the 10% size. No task-F1 crossover appears over the evaluated training sizes. The $\lambda \propto 1/n$ schedule of Theorem 4 does not transfer either. Scaling the penalty by that rule loses to the fixed full-size weight by 0.46 F1 at 25% and 0.22 F1 at 50%, on three of three seeds each. At 10% its mean is 0.30 F1 ahead, but on one seed of three and through a single 2.01 F1 outlier. The optimized linear-model prescription therefore does not yield a consistent F1 improvement under these pretrained-model budgets.

D.4 BINARY-RULE PRECISION ON THE GOLD POOLS

For a binary selector we use a single nonempty selection threshold with $T = 1$ and $\alpha = \delta = .05$. With t valid and f invalid selected coordinates, the lower bound is $L_\delta^- = \mathrm{Beta}^{-1}(\delta, t, f + 1)$. Table 13 reports the resulting coordinate-binomial calculation. Both lower bounds exceed the numerical

Table 12: Data-size scan on DocRED. The ΔCVR rows are seed-paired means with the number of same-sign seeds in parentheses. Every size uses the three seeds common to the whole scan, including full size. The four columns are therefore internally comparable. The five-seed full-size contrasts appear separately in Table 2. The F1 rows are method means over three seeds at the subsampled sizes and 5 at full size. The CVR contrasts increase in all three seeds. F1 uses the explicitly stated run counts and is a separate outcome.

	10%	25%	50%	100%
Training documents	305	763	1,526	3,053
Collateral violation rate, ΔCVR (paired)				
Indiscrim. − ERM	.022 (3/3)	.093 (3/3)	.138 (3/3)	.192 (3/3)
Random − validated	.020 (3/3)	.091 (3/3)	.133 (3/3)	.194 (3/3)
Development F1 (means, $n = 3, 3, 3, 5$ runs)				
ERM	52.26	56.22	58.95	61.01
Indiscriminate	52.02	55.98	58.54	60.66
Random mask	51.81	56.21	58.61	60.89
VCC (R-VAL)	51.62	56.51	58.92	60.90
VCC, $\lambda \propto 1/n$	51.92	56.05	58.70	–

precision target $1 - \alpha = .95$. The pooled coordinates are correlated within sentences or documents and DocRED uses the same development split for model selection. These summaries do not establish the independent calibration sampling required for an unconditional population guarantee under Theorem 6.

Table 13: Descriptive binary-rule precision on the gold pools. The last column is a one-sided coordinate-binomial Clopper–Pearson calculation at $\delta = .05$. Coordinates within documents or sentences are correlated, and development data also select the model. These bounds do not establish independent-calibration population guarantees.

Domain	Valid selected	Invalid selected	Precision	Binomial bound
DocRED	7,411	178	.977	.973
CoNLL-2003	5,792	165	.972	.969

D.5 Prediction agreement on preserved DocRED coordinates

Table 14 complements the altered-evidence measurements with the same near-invariance statistic on preserved gold coordinates. At the primary tolerance the six method means lie in $[.970, .981]$ with band 1.2 percentage points. This concentration does not imply zero differences or statistical equivalence. The rates measure model responses and are neither task accuracy nor a test that conditional label probabilities remain unchanged.

Table 14: DocRED near-invariance rates on preserved gold coordinates. Each model uses $9{,}019$ coordinates. Entries are means (standard deviations) over five seeds.

Method	$\kappa = .05$	$\kappa = .1$	$\kappa = .2$
ERM	.955 (.001)	.970 (.001)	.982 (.001)
Indiscriminate	.950 (.002)	.976 (.001)	.991 (.000)
SIEF-style	.963 (.003)	.981 (.001)	.992 (.001)
Random mask	.949 (.005)	.976 (.002)	.991 (.001)
VCC (R-VAL)	.941 (.002)	.970 (.002)	.987 (.001)
Oracle	.943 (.004)	.972 (.003)	.988 (.002)

E CoNLL-2003 Settings and Full Tables

E.1 Transplanting the evaluation design

This section specifies the CoNLL data and optimization choices used in Section 6.2.

The CoNLL-2003 measurement is the DocRED measurement with each object replaced by its named-entity counterpart, with the pool, backbone and optimization specified below.

Coordinate. A word position j in a sentence. The gold pool is all $(j, c_j^\star)$ with $c_j^\star \neq \texttt{O}$ over the augmentable development sentences, 8,215 coordinates in total.

Augmentation. Replace a Bernoulli(0.15) subset of the words of a sentence by `[MASK]`. The word count is preserved, so the two views align one-to-one by word index. Sentences shorter than four words are not augmented, mirroring DocRED's treatment of documents shorter than four sentences. 2,874 of the 3,250 development sentences qualify.

Validity. $V_j = 0$ when j carries a gold entity label and at least one word of the gold mention containing j has been masked, so that the surface evidence for that label has been destroyed. This is the counterpart of DocRED's "the deleted sentence belongs to the evidence set of the triple". Of the 2,236 violated development coordinates, 1,213 are violated because the token itself was masked. Another word of the mention was masked in 870 cases. Both reasons apply to 153 coordinates.

Active pool. $\mathcal{K}(x) = \{j : 1 - p_{\text{ref}}(x)[j, \texttt{O}] > 0.01\}$, computed once on the clean view with a fixed reference tagger, which is then released. It costs no forward evaluations during training. The active fraction is .167 of all coordinates.

Rule validator. Position j is declared invalid when j is masked, or when j lies in the same maximal capitalised run as a masked word. The rule reads no labels. On the development gold pool it attains precision .972, recall .969 and selection rate .725. On the training pool, .977, .965 and .717. The bookkeeping identity $ps = r(1 - \pi)$ of (17) holds on both pools, at .7051 on development and .7010 on training.

Random-mask rate. The random-mask method draws coordinates independently at rate .715, the rule validator's measured selection rate over the active training pool, so that the two masks have comparable realized selection fractions.

Metric. $\text{CVR}_\kappa = \mathbb{P}[|p_\theta(x)[j, c^\star] - p_\theta(Ax)[j, c^\star]| < \kappa \mid V_j = 0]$ for $\kappa \in \{0.05, 0.1, 0.2\}$, primary 0.1, reported as an integer hit count over the integer denominator 2,236.

Backbone and optimization. We use `bert-base-cased` token classification over 9 BIO labels. The learning rate is 5×10^{-5} for the encoder and 10^{-4} for the head. We train for two epochs at batch 48 with maximum length 128. Warmup is 0.06 and gradient clipping is 1.0. Training uses bf16 autocast and every reported number uses fp32 softmax. Model selection is the best development entity-F1 epoch, and CVR is reported at that same epoch. All methods use the same two-epoch budget and batch size.

Reference tagger. The fixed tagger f_{ref} that fixes the active pool is a separate ERM model at seed 0, trained for three epochs at batch 32 under the same optimizer settings. It supplies only the active pool and is separate from the ERM models used in the method comparison.

Consistency term. Symmetric KL between the two views' per-position softmax, averaged over the selected coordinates only, at $\lambda = 1.0$.

Determinism of the measurement. Development is evaluated in a fixed length-sorted order at augmentation seed 12345, so the gold pool, the violated set and the rule verifier's operating point are identical across every method and seed. Training uses length bucketing inside a shuffled window of 50 batches on one device.

E.2 Method table and per-seed values

This section expands the CoNLL block of Table 2 with the complete task and response measurements.

E.3 The data-size scan

Holding the epoch budget fixed also changes the number of optimizer updates as the training subset grows. The size scan therefore describes fitted models under a common training rule.

Table 17 reports three matched initializations at each training size with batch 48 and a two-epoch budget. Mean indiscriminate-minus-ERM collateral excess decreases across the first three sizes and is approximately flat from half to full size. Table 16 separates the final increment from the full-range change. Both intervals span zero. Positive excess at each size supports persistence, while the scan does not establish a monotone population trend.

Table 15: CoNLL-2003 at full training size, every method at every seed. Entity F1 and the collateral violation rate at $\kappa \in \{.05, .1, .2\}$. The collateral columns are integer hit counts over the fixed denominator 2,236 of violated gold coordinates, identical for every run. "Pen. frac." is the realised fraction of active coordinates the method penalises. The last column gives the number of seeds behind each method mean.

Method	F1	$\text{CVR}_{.05}$	$\text{CVR}_{.1}$	$\text{CVR}_{.2}$	Pen. frac.	Seeds
ERM (backbone)	93.93 (0.36)	.421	.487	.547	–	3
Indiscriminate (UDA)	90.99 (0.07)	.610	.718	.832	1.000	3
Random mask	90.65 (0.45)	.594	.714	.829	.716	3
VCC (R-VAL)	91.97 (0.30)	.295	.382	.476	.716	3
Per-seed development F1 and $\text{CVR}_{.1}$ hit count out of 2,236						
ERM (backbone)	66 (94.13 / 1111)	67 (94.14 / 1064)	68 (93.51 / 1091)			
Indiscriminate (UDA)	66 (90.98 / 1595)	67 (91.06 / 1613)	68 (90.93 / 1606)			
Random mask	66 (90.27 / 1573)	67 (91.14 / 1610)	68 (90.54 / 1608)			
VCC (R-VAL)	66 (92.04 / 888)	67 (91.64 / 811)	68 (92.23 / 861)			

Table 16: Changes in all-output-minus-ERM CVR excess on CoNLL, in percentage points. Mean changes and paired t intervals use three shared seeds. The sign-flip test enumerates all eight assignments.

Training-share contrast	Mean	95% interval	Sign-flip p
$100\% - 50\%$	$+0.03$	$[-14.57, 14.63]$	1.00
$100\% - 12.5\%$	-2.22	$[-12.97, 8.52]$	–

E.4 Pool-side comparison with DocRED

This section supports the comparison of validator operating points in Section 6.3 with the domain-level pool quantities.

E.5 Sensitivity on preserved coordinates

This section supports Section 6.5 by comparing prediction agreement on preserved and altered evidence.

The preserved gold pool contains 5,979 coordinates. At $\kappa = .1$ the method means range from .958 to .971 with a band of 1.35 percentage points. The collateral rates at all evaluated tolerances appear in Table 15. For altered coordinates, CVR under all-output and random penalties remains above ERM while the rule's CVR remains below it. Figure 2 displays both coordinate groups at the primary tolerance.

F Extended Related Work

Table 19 locates the contribution against the families surveyed in Section 2. "Selection" is the unit at which the penalty can be switched off. "Coord." marks a selector resolved per output coordinate rather than per example. "Dual" marks a method that also uses the coordinates it judges violated as supervision. "Certified" marks a finite-sample guarantee on the correctness of the selector.

Two entries deserve a longer reading. SIEF (Xu et al., 2022) already selects its consistency term per relation coordinate, keeping only coordinates whose cross-view score shift is small, so the need for coordinate selection has been felt in practice. Its selector, however, is scored by the model being trained. Proposition 11(ii) shows that this circular estimate inherits the error of the problem the penalty addresses, and it carries no guarantee on how often it is right. Appendix D measures what happens to it under its own training. Loss-correction methods for noisy labels (Natarajan et al., 2013; Patrini et al., 2017) use class-conditional noise rates or a noise transition matrix, whereas we calibrate validity within the selected subset directly. The noise in our setting is also structured rather than class-conditional, landing precisely on the coordinates the augmentation disturbed, which are the coordinates on which the augmentation carries the most information.

Table 17: CoNLL-2003 data-size scan with three matched seeds at every size and the evaluation design of Table 15. The excess collateral rate of indiscriminate consistency over ERM remains positive at every training size without the growth pattern observed on DocRED.

Share	Sentences	Entity F1		$CVR_{.1}$		Excess
		ERM	UDA	ERM	UDA	UDA−ERM
12.5%	1,755	70.85	63.35	.431	.684	+.253
25%	3,510	87.16	80.90	.408	.647	+.239
50%	7,020	92.07	87.64	.446	.677	+.230
100%	14,041	93.93	90.99	.487	.718	+.231

Table 18: Pool-side quantities on the two domains, measured before any method is trained. Both selection rates and precisions are read on the annotated gold pool. The rows distinguish the gold pool from the full coordinate population. The conflict floor is $\frac{1}{4}\pi\gamma^2$ under the exact-preservation convention $\gamma^2{=}1$, which is stated and not measured. What is measured on each domain is π.

	DocRED	CoNLL-2003
Violation mass $\hat{\pi}$ on the gold pool	.258	.272
Violation mass over all coordinates	.208	.177
Conflict floor $\frac{1}{4}\hat{\pi}\gamma^2$	.0646	.0680
Rule-validator selection rate	.624	.725
Rule-validator precision	.977	.972

Within relation extraction specifically, entity-replacement and counterfactual studies test and enforce invariance to renaming (Wang et al., 2023b; Modarressi et al., 2024), and self-distillation enforces agreement between relational reasoning views (Zhang et al., 2023). Augmentation-consistency terms also regularize weakly and complementarily labelled multi-label objectives (Gao et al., 2026). These objectives share the property the analysis isolates. The penalty is issued for every output coordinate of a view at once.

Prediction targets The Pi-model matches stochastic predictions (Laine & Aila, 2017) and Mean Teacher averages model weights to stabilize its targets (Tarvainen & Valpola, 2017). UDA matches predictions under stronger transformations (Xie et al., 2020).

Perturbation choices FixMatch selects confident targets from weak augmentations for learning on strong augmentations (Sohn et al., 2020), while virtual adversarial training selects locally adversarial perturbations (Miyato et al., 2019) and R-Drop matches distributions across dropout realizations (Wu et al., 2021; Li et al., 2024a).

Environmental invariance Invariant-risk objectives seek predictors shared across environments (Arjovsky et al., 2019), subject to conditions that determine whether environmental invariance transfers (Rosenfeld et al., 2021).

Evidence and altered targets. Explanatory spans (Lei et al., 2016) and counterfactually edited examples (Kaushik et al., 2020) locate information supporting a target. Negative augmentation uses structurally corrupted views as negative examples (Sinha et al., 2021), whereas counterfactual invariance suppresses nuisance attributes while retaining relevant information (Veitch et al., 2021). Loss correction uses class-conditional noise transitions (Natarajan et al., 2013; Patrini et al., 2017). Here the augmentation locates potentially altered outputs and independent validity labels assess whether each selected target survives, connecting evidence to the reliability of penalty placement within an example.

Calibration extensions Conformal p-values (Jin & Candes, 2023) and selective labeling (Huang et al., 2025) control errors among selected items, while prediction-powered inference (Einbinder et al., 2025) uses unlabeled data for calibration and conformal risk training (Yeh et al., 2025) optimizes the predictor for risk control.

Selection and pseudo-label quality. Co-teaching exchanges small-loss examples between two networks (Han et al., 2018). DivideMix separates examples by modeled loss and reuses the uncertain

Table 19: Where the validity assumption enters.

Family	Selection	Coord.	Dual	Certified
Π-model, Mean Teacher	none	–	–	–
VAT, R-Drop	none	–	–	–
UDA confidence	example	–	–	–
FixMatch confidence	example	–	–	–
Self-distillation	none	–	–	–
SIEF	self-scored coord.	yes	–	–
Class-aware SSL threshold	self-scored class	partial	–	–
Loss correction	class prior	partial	–	–
VCC (this work)	certified coord.	yes	yes	yes

subset through semi-supervised learning (Li et al., 2020). These approaches address corrupted training annotations (Song et al., 2023). Pseudo-labeling can reinforce incorrect predictions, which motivates regularization and balanced use of labeled examples (Arazo et al., 2020). FreeMatch adapts confidence thresholds to training progress and class-level predictions (Wang et al., 2023a). We assess augmentation-induced target changes at individual outputs rather than the correctness of a predicted training label.

Calibrated supervision. Conformal credal self-supervision calibrates sets of label distributions and trains against their closest admissible targets (Lienen et al., 2023). Reliable programmatic weak supervision gives confidence intervals for group label probabilities from labeling-function constraints (Álvarez et al., 2025). Conformal risk control bounds expected monotone loss (Angelopoulos et al., 2024). PAC labeling combines predicted and expert labels to bound error over a fixed dataset with high probability (Candès et al., 2025). Our calibration target is the fraction of augmentation-invalid outputs among those receiving a consistency penalty.

Local evidence and annotation granularity. ReLabel supplies spatial label maps so that cropped images can receive localized supervision (Yun et al., 2021). Its targets come from a pretrained teacher, whereas our validity labels assess whether a transformation preserves an output's supporting evidence. SQuAD 2.0 introduces unanswerable questions with plausible distractors (Rajpurkar et al., 2018). This benchmark underlies the supplementary question-answering comparison in Appendix B.4.