DOES CERTIFIED ROBUSTNESS GENERALIZE? A NEURAL NETWORK VERIFICATION STUDY OF DISTRIBUTION SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Certified robust accuracy (CRA) is the standard metric for quantifying worst-case adversarial robustness and for benchmarking certified defenses. However, CRA is by construction reported only on the in-distribution (ID) test set on which the model was trained. A largely understudied question is what CRA tells us about broader model behavior, in particular predictive accuracy, and whether the relationship between certified robustness and accuracy carries a transferable structure under distribution shift. This is non-trivial because formal verification is strictly stronger than any single-attack evaluation: a CRA guarantee asserts the absence of any adversarial example in an ℓ_∞-ball, so the "on-the-line" transfer observed for clean accuracy and empirical attacks need not carry over. We present, to the best of our knowledge, the first systematic study of CRA under distribution shift for verification-tractable architectures, covering MNIST-FC, GTSRB-CNN, and CIFAR-CNN across structural and corruption shifts. Using clean-accuracy-matched model families and shared-property verification, we find that (i) the inter-model CRA ordering shows strong but imperfect agreement from ID to OOD, with statistically significant ID advantages retained under the tested shifts, and (ii) ID CRA correlates with OOD accuracy, calibration, and OOD detection across a multi-architecture ablation, acting as a model-selection proxy for several practical OOD properties. We further support these findings with a machine-checked theoretical framework that characterizes when ID CRA bounds OOD performance and with controlled-shift experiments that test the bound directly. Selecting by ID CRA incurs 0–7 percentage points of OOD CRA loss in the tested settings. These findings extend CRA's value beyond robustness benchmarking, supporting its use as a model-selection signal for accuracy, calibration, and detection performance under distribution shift. Our code is available at `https://anonymous.4open.science/r/veri-generalization-anonymous`.

1 INTRODUCTION

Certified robustness verification provides a formal guarantee that no input within an ℓ_∞-ball of radius ε around a test point changes the model's prediction. Certified Robust Accuracy (CRA) aggregates these per-instance guarantees and is the standard metric for benchmarking certified defenses (Li et al., 2023; Croce et al., 2021; Johnson, 2025). Modern verification methods span SMT, bound-propagation, mixed-integer, and smoothing-based approaches (Katz et al., 2017; Tjeng et al., 2019; Wang et al., 2021; Cohen et al., 2019), and increasingly support convolutional and residual networks (Bak et al., 2021; Zhang et al., 2022).

CRA is nevertheless computed on a fixed in-distribution (ID) test set, although deployed models encounter shifted inputs caused by changes in sensors, lighting, geography, and acquisition conditions (Koh et al., 2021; Recht et al., 2019; Hendrycks & Dietterich, 2019; Hendrycks et al., 2021a). Such shifts can produce silent failures in medical imaging and autonomous driving (DeGrave et al., 2021; Filos et al., 2020; Hendrycks et al., 2021b). If certified guarantees collapse or defense rankings invert under shift, ID-only evaluation gives a misleading picture of deployment robustness.

Despite this practical concern, it remains unclear how much an ID certificate reveals about robustness and predictive behavior on out-of-distribution (OOD) inputs. Prior work studies the transfer of

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

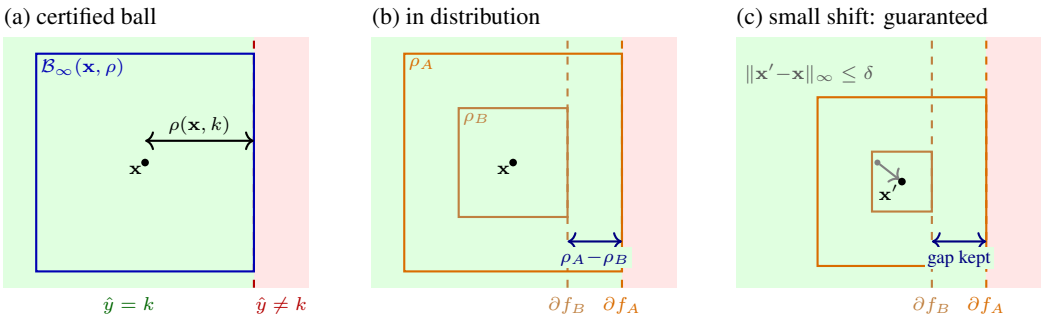


Figure 1: Theorem 1 in pictures. **(a)** The certified radius $\rho(\mathbf{x}, k)$ is the half-side of the largest ℓ_∞ ball around $\mathbf{x}$ on which the prediction stays k. **(b)** In distribution, the decision boundary of f_A is farther from $\mathbf{x}$ than that of f_B, so $\rho_A > \rho_B$; over a test set this yields a CRA gap $\Delta_{\mathcal{S}_P} > 0$. **(c)** A shift moving $\mathbf{x}$ by at most δ shrinks each certified radius by at most δ (Proposition 1), so the gap erodes by at most $s \leq \alpha_A^P + \alpha_B^P + 2\hat{\eta}(\delta)$; the order is guaranteed when $\Delta_{\mathcal{S}_P}$ exceeds this cap.

clean-accuracy rankings and attack-based robustness under natural shift (Taori et al., 2020; Miller et al., 2021; Croce & Hein, 2020); certified robustness is different because its guarantee covers every perturbation in a neighborhood and its reported value depends on verifier bounds rather than one attack. Re-verifying every model on every OOD slice is also costly, motivating ID CRA as a potential model-selection proxy for shifted performance.

This paper. We conduct, to the best of our knowledge, the first systematic study of how certified robust accuracy behaves under distribution shift for verification-tractable architectures. Our primary question is: *given evidence for in-distribution certified robustness, how much can we infer about a model's robustness and predictive behavior on out-of-distribution inputs?* We address this question by developing theoretical bounds and testing their implications empirically. Under a label-preserving coupling, our first theorem bounds the reduction in the CRA gap between two models using their ID CRA curves and a shift-defect term, the fraction of coupled inputs displaced beyond a specified distance. The resulting sufficient condition guarantees that the ordering is preserved whenever the ID advantage exceeds this bound (Theorem 1, Corollary 1; Figure 1). Our second theorem bounds OOD prediction-induced loss in terms of ID CRA and the same defect, yielding a lower bound on OOD classification accuracy (Theorem 2, Corollary 2). These results, formalized in Lean 4, identify the model- and shift-dependent quantities that govern the guarantees.

We design three complementary experiments around this framework: *(i) Natural-shift transfer.* We compare CRA rankings within clean-accuracy-matched MNIST-FC, GTSRB-CNN, and CIFAR-CNN families under structural and corruption shifts, using shared property sets and paired comparisons to measure changes in the ID CRA gaps against that bound (Section 5.1; Figure 2). *(ii) Controlled-shift validation.* We vary displacement and defect by construction to evaluate the transfer bounds, quantify their tightness, and test the predicted ordering when the sufficient condition holds (Section 5.2). *(iii) Model-selection utility.* Motivated by the loss bound, we examine whether ID CRA tracks OOD accuracy and broader probabilistic metrics, including calibration and detection, across a 17-model architecture ablation. We also measure the OOD CRA lost by selecting a model using ID CRA rather than OOD evaluation (Section 5.3). Together, these experiments assess both the quantitative guarantees and the empirical usefulness of ID CRA under distribution shift.

Contributions. Our contributions are as follows.

- We introduce the first systematic framework for studying how in-distribution CRA relates to out-of-distribution behavior, combining matched-clean-accuracy model families on three datasets, shared VNNLIB property sets per distribution, a controlled-shift axis with known transport defect, a 17-model architecture ablation, a three-verifier portfolio (α, β-CROWN, Marabou, and NNV) for incompleteness-robust CRA estimates, and a six-metric OOD evaluation panel, so that observed CRA differences reflect real robustness differences rather than evaluation noise or single-verifier artifacts.
- We establish, both theoretically and empirically, two transfer results: certified-robustness rankings show strong but imperfect agreement from ID to both structural and corruption OOD (ρ from 0.77 to 0.96 across the architecture ablation and the CIFAR family), with paired tests separating plain from robust models; and ID CRA itself transfers as a predictive signal for several OOD properties,

correlating with OOD probabilistic fit ($\rho \leq -0.89$ on NLL), classification accuracy, and OOD detection, with the cost of selecting a model by ID CRA measured rather than assumed.

- We draw out the broader implications for neural-network verification: ID CRA can serve as a cheap deployment-time proxy for selecting models under shift; adversarial training delivers most of its OOD value through calibration rather than accuracy; and benchmarking practice should account for query-level input-difficulty differences, as illustrated by our GTSRB anomaly in which every model's structural-OOD CRA exceeds its ID CRA.

All claims are made for the verification-tractable regime, the networks on which complete ℓ_∞ verification is currently feasible; ImageNet-scale models lie outside it by the nature of the tool, not by choice.

2 Related Work

Neural-network verification. SMT-based methods (Katz et al., 2017), abstract interpretation (Gehr et al., 2018) and mixed-integer programming (Tjeng et al., 2019) give complete or near-complete reasoning at significant cost; bound propagation (CROWN (Zhang et al., 2018), α-CROWN (Xu et al., 2021), β-CROWN (Wang et al., 2021)) underlies α, β-CROWN (Zhang et al., 2022), the VNN-COMP winner (Bak et al., 2021) and our verification backbone. Every one of them reports CRA on a fixed in-distribution test set.

Adversarial training and certified training. PGD adversarial training (Madry et al., 2018) is the dominant empirical defense; TRADES (Zhang et al., 2019) adds a KL regularizer that controls the clean–robust trade-off. Certified training such as IBP and CROWN-IBP (Gowal et al., 2019; Mirman et al., 2018) optimizes a verifier-aware loss, improving CRA at the cost of clean accuracy and of coupling to one verifier family. We use PGD-AT and TRADES so that CRA differences between models come from the training objective and $\varepsilon_{\text{train}}$, not from verifier-specific losses.

OOD generalization, calibration, and shift datasets. Clean-accuracy rankings of image classifiers are remarkably stable under natural distribution shift (Taori et al., 2020), formalized as "accuracy on the line" by Miller et al. (2021), and empirical adversarial robustness under attacks like AutoAttack (Croce & Hein, 2020) on benchmarks such as ImageNet-C (Hendrycks & Dietterich, 2019) largely preserves ID orderings. Networks are miscalibrated on clean data (Guo et al., 2017) and more so under shift (Ovadia et al., 2019), motivating ECE, Brier, NLL and the max-softmax OOD baseline (Hendrycks & Gimpel, 2017); adversarial training acts as an implicit confidence regularizer on shifted inputs (Augustin et al., 2020). Whether either regularity carries over to certified robustness is what this paper measures, using EMNIST (Cohen et al., 2017), the GTSRB (Stallkamp et al., 2012) / BTSD (Timofte et al., 2014) pair, PixMix (Hendrycks et al., 2022) and CIFAR-10-C (Hendrycks & Dietterich, 2019).

Certificate construction versus certificate transfer. Randomized smoothing (Cohen et al., 2019) and sample-based certification (Anderson & Sojoudi, 2022) produce probabilistic guarantees on a noise-perturbed input distribution, and recent defenses such as HyCAS (Dhar et al., 2026) improve the radius–noise trade-off of such certificates by combining Lipschitz-constrained convolutions with stochastic components. Those works improve how a certificate is *computed*. This paper takes the certificate as given and asks whether its in-distribution value *transfers*; Appendix F.2 instantiates it for smoothing certificates. Domain-adaptation, PAC-Bayes and distributionally robust bounds are related to Theorem 2; Appendix A.9 places it among them.

3 Background and Theoretical Framework

Theorem 1 gives a sufficient condition, in quantities measurable from in-distribution evaluation, under which the ranking of two classifiers by CRA is preserved on a shifted distribution; Theorem 2 bounds OOD loss by ID CRA. Proofs and the Lean 4 formalization are in Appendix A.

3.1 Definitions

Neural network setups. Let $f_\theta : \mathbb{R}^d \to \mathbb{R}^K$ be a feedforward network with parameters θ, mapping inputs $\mathbf{x} \in [0,1]^d$ to K logits over the label space $\mathcal{Y} = \{1, \ldots, K\}$, with predicted label $\hat{y}(\mathbf{x}) = \arg\max_j f_j(\mathbf{x})$.

Definition 1 (Local ℓ_∞ robustness)**.** *f_θ is* locally robust *to input x with true class k at radius $\varepsilon \in \mathbb{R}$ if $\hat{y}(\mathbf{x}') = k$ for every $\mathbf{x}' \in \mathcal{B}_\infty(\mathbf{x}, \varepsilon) := \{\mathbf{x}' \in [0,1]^d : \|\mathbf{x}' - \mathbf{x}\|_\infty \le \varepsilon\}$.*

Definition 2 (Certified Robust Accuracy)**.** *For a network f_θ, test set $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, radius ε,*

$$\mathrm{CRA}(f_\theta, \mathcal{S}, \varepsilon) \;=\; \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\big[f_\theta \text{ is verified robust at } (\mathbf{x}_i, y_i) \text{ at radius } \varepsilon\big], \tag{1}$$

where $\mathbf{1}[\cdot]$ denotes the indicator function, taking value 1 when its argument holds and 0 otherwise; instances on which the verifier returns SAT *or* unknown *are counted as not verified.*

Distribution Shift Taxonomy. Let P be the source distribution and Q a shifted distribution on $[0,1]^d \times \mathcal{Y}$ with the same label space. We consider **in-distribution** ($Q = P$); **structural OOD**, where Q arises from a different data-generating process with the same classes (a different handwriting source, a different country's signs); and **corruption OOD**, where Q applies a stochastic corruption to samples from P (PixMix (Hendrycks et al., 2022)). Both shifts preserve the label, so CRA is well-defined on Q.

3.2 Theoretical Framework

The Certified Radius and Its 1-Lipschitz Property. The theoretical framework is built on the following property of the certified radius.

Definition 3 (Certified ℓ_∞ radius)**.** *The certified ℓ_∞ radius of f_θ at $(\mathbf{x}, k)$ is*

$$\rho_{f_\theta}(\mathbf{x}, k) \;:=\; \sup\big\{\, \varepsilon \ge 0 : f_\theta \text{ is locally robust at } (\mathbf{x}, k) \text{ at radius } \varepsilon \,\big\}, \tag{2}$$

with $\rho_{f_\theta}(\mathbf{x}, k) = 0$ when $\hat{y}(\mathbf{x}) \ne k$.

Every result below concerns true local robustness, which the radius brackets: $\rho_{f_\theta}(\mathbf{x}, k) > \varepsilon$ implies local robustness at radius ε, which implies $\rho_{f_\theta}(\mathbf{x}, k) \ge \varepsilon$. A sound verifier reports a lower bound $\hat{\rho}_{f_\theta} \le \rho_{f_\theta}$ (Appendix B): a `safe` verdict at $(\mathbf{x}, \varepsilon + \delta)$ certifies true robustness at radius ε for every $\mathbf{x}'$ with $\|\mathbf{x}' - \mathbf{x}\|_\infty \le \delta$, but it does not guarantee that a separate verifier run at $\mathbf{x}'$ returns `safe`. Comparisons between reported CRA values in Section 5 are therefore empirical checks of the bounds; they are exact in cells where the verifier decided every instance, and we report the number of undecided instances per cell. Smoothing certificates (Section 5) are Monte-Carlo estimates that hold with probability $1 - \alpha$ per input, so a statement over M inputs holds with probability at least $1 - M\alpha$.

Proposition 1 (1-Lipschitz certified radius)**.** *For every classifier f_θ, every label $k \in \mathcal{Y}$, and every $\mathbf{x}, \mathbf{x}' \in [0,1]^d$,*

$$\big|\rho_{f_\theta}(\mathbf{x}, k) - \rho_{f_\theta}(\mathbf{x}', k)\big| \;\le\; \|\mathbf{x} - \mathbf{x}'\|_\infty. \tag{3}$$

Proof. Set $\delta := \|\mathbf{x} - \mathbf{x}'\|_\infty$ and $r := \rho_{f_\theta}(\mathbf{x}, k)$, and write $\mathcal{B}_\infty(\mathbf{u}, s) := \{\mathbf{v} : \|\mathbf{v} - \mathbf{u}\|_\infty \le s\}$ for the closed ℓ_∞-ball of radius s around $\mathbf{u}$. For any $r_0 < r$ and any $\mathbf{z}' \in \mathcal{B}_\infty(\mathbf{x}', r_0 - \delta)$ (vacuous if $r_0 \le \delta$), the triangle inequality gives $\|\mathbf{z}' - \mathbf{x}\|_\infty \le (r_0 - \delta) + \delta = r_0$, so $\mathbf{z}' \in \mathcal{B}_\infty(\mathbf{x}, r_0)$ and $\hat{y}(\mathbf{z}') = k$ by Definition 1. Taking $r_0 \to r$ yields $\rho_{f_\theta}(\mathbf{x}', k) \ge r - \delta$. The other direction follows by symmetry. □

The containment step of this proof, that local robustness at $(\mathbf{x}, k)$ at radius $\varepsilon + \delta$ implies it at $(\mathbf{x}', k)$ at radius ε whenever $\|\mathbf{x}' - \mathbf{x}\|_\infty \le \delta$, is the only property of f_θ used below; it relies solely on Definition 1 and the triangle inequality on ℓ_∞, with no continuity, smoothness, or architectural assumption on f_θ. The argument uses only the triangle inequality, so every result below holds for any norm (the Lean development is over an arbitrary normed space); we fix ℓ_∞ because it is what verifiers certify, and report the ℓ_2 instance via smoothing in Section 5.

Rank Preservation under Shift. We now formulate the shift assumptions and state the main result. A *coupling* of (P, Q) is a joint distribution on $([0,1]^d \times \mathcal{Y})^2$ with marginals P and Q.

Assumption 1 (Label-preserving coupling)**.** There exists a coupling π of (P, Q) such that for $((\mathbf{x}, Y), (\mathbf{x}', Y')) \sim \pi$, $Y = Y'$ holds π-almost surely.

Assumption 1 is a condition on the label marginals only: a label-preserving coupling of (P, Q) exists if and only if $P_Y = Q_Y$ (Lemma 1, Appendix A.2). It does not require a sample-wise correspondence between the domains. Every shift axis in Section 4 satisfies it by construction: property sets are class-balanced, corruptions preserve labels, and the BTSD subset is class-balanced. When the label marginals differ, the same lemma shows that every bound below holds with $\hat{\eta}(\delta)$ replaced by $\hat{\eta}(\delta) + \mathrm{TV}(P_Y, Q_Y)$.

Lemma 1 (Label-preserving couplings)**.** *Let P_Y, Q_Y be the label marginals of P and Q.* (i) *A coupling π of (P, Q) with $Y = Y'$ π-a.s. exists if and only if $P_Y = Q_Y$.* (ii) *In general there is a coupling π of (P, Q) with $\pi(Y \neq Y') = \mathrm{TV}(P_Y, Q_Y)$, and under any coupling every bound below holds with $\hat{\eta}(\delta)$ replaced by the fraction of indices i with $\|\mathbf{x}'_i - \mathbf{x}_i\|_\infty > \delta$ or $y_i \neq y'_i$, which is at most $\hat{\eta}(\delta) + \widehat{\mathrm{TV}}$, the second term being the fraction of label-changing indices.* (Proof in Appendix A.2.)

Assumption 2 (Bounded ℓ_∞ transport)**.** *For some $\delta \geq 0$, the coupling π has shift defect $\eta_\pi(\delta) := \pi\big(\{(\mathbf{x}, \mathbf{x}') : \|\mathbf{x}' - \mathbf{x}\|_\infty > \delta\}\big) \in [0, 1]$.*

For index-aligned coupled samples $\mathcal{S}_P = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ and $\mathcal{S}_Q = \{(\mathbf{x}'_i, y_i)\}_{i=1}^N$ drawn from π, the empirical shift defect is

$$\hat{\eta}(\delta) \;:=\; \frac{1}{N}\,\big|\{\, i : \|\mathbf{x}'_i - \mathbf{x}_i\|_\infty > \delta \,\}\big|. \tag{4}$$

Assumption 2 restricts nothing: for any coupling and δ it defines the defect $\hat{\eta}(\delta)$, so the results below hold for every coupling and every $\delta \geq 0$; the substantive question is where the bound is informative (Appendix A.8), which Section 5.2 settles by measuring $\hat{\eta}$.
The single-model form of the transfer bound is what the controlled experiments of Section 5.2 test directly.

Lemma 2 (CRA shift bound)**.** *Under Assumptions 1–2, for any classifier f_θ, every $\varepsilon \geq 0$, and every $\delta \geq 0$,*

$$\mathrm{CRA}(f_\theta, \mathcal{S}_Q, \varepsilon) \;\geq\; \mathrm{CRA}(f_\theta, \mathcal{S}_P, \varepsilon + \delta) \;-\; \hat{\eta}(\delta), \tag{5}$$

$$\mathrm{CRA}(f_\theta, \mathcal{S}_Q, \varepsilon + \delta) \;\leq\; \mathrm{CRA}(f_\theta, \mathcal{S}_P, \varepsilon) \;+\; \hat{\eta}(\delta). \tag{6}$$

The proof (Appendix A.1) applies Proposition 1 to each coupled pair and charges the indices displaced by more than δ to the defect.
Given two classifiers f_A, f_B, write $\Delta_{\mathcal{S}}(\varepsilon) := \mathrm{CRA}(f_A, \mathcal{S}, \varepsilon) - \mathrm{CRA}(f_B, \mathcal{S}, \varepsilon)$ for CRA gap on $\mathcal{S}$.

Theorem 1 (Pairwise rank preservation)**.** *Under Assumptions 1–2, for every $\varepsilon \geq 0$ and every $\delta \geq 0$,*

$$\Delta_{\mathcal{S}_Q}(\varepsilon) \;\geq\; \Delta_{\mathcal{S}_P}(\varepsilon) \;-\; \alpha_A^P(\varepsilon, \delta) \;-\; \alpha_B^P(\varepsilon, \delta) \;-\; 2\hat{\eta}(\delta), \tag{7}$$

where the (non-negative) modulus terms are

$$\alpha_A^P(\varepsilon, \delta) := \mathrm{CRA}(f_A, \mathcal{S}_P, \varepsilon) - \mathrm{CRA}(f_A, \mathcal{S}_P, \varepsilon + \delta), \tag{8}$$

$$\alpha_B^P(\varepsilon, \delta) := \mathrm{CRA}(f_B, \mathcal{S}_P, \varepsilon - \delta) - \mathrm{CRA}(f_B, \mathcal{S}_P, \varepsilon). \tag{9}$$

Proof sketch (full proof in Appendix A.3; Lean-checked, Appendix A.5). Apply (5) to f_A and (6) to f_B at radius $\varepsilon - \delta$ and subtract; no concentration is needed. Pointwise, the same argument gives $\mathrm{CRA}(f_\theta, \mathcal{S}_Q, \varepsilon) \geq N^{-1} \sum_i \mathbf{1}[\rho_{f_\theta}(\mathbf{x}_i, y_i) - \|\mathbf{x}'_i - \mathbf{x}_i\|_\infty > \varepsilon]$ (Remark 1), with each pair's own displacement in place of δ.

Corollary 1 (Sufficient condition for rank preservation)**.** *If*

$$\Delta_{\mathcal{S}_P}(\varepsilon) \;>\; \alpha_A^P(\varepsilon, \delta) + \alpha_B^P(\varepsilon, \delta) + 2\hat{\eta}(\delta), \tag{10}$$

then $\Delta_{\mathcal{S}_Q}(\varepsilon) > 0$, and the in-distribution ranking of f_A and f_B at radius ε is preserved on $\mathcal{S}_Q$.

Every quantity in (10) is observable from in-distribution evaluation plus a coupling, so the condition is checkable *a priori*: the theorem is a deterministic statement machine-checked in Lean 4 (Appendix A.5), and the testable predicate is the budget condition. The bound is two-sided (Appendix A.4), so a Q with higher absolute CRA than P is consistent with it provided the order is preserved, as on the BTSD axis. Equivalently, with erosion $s := \Delta_{\mathcal{S}_P} - \Delta_{\mathcal{S}_Q}$ (the CRA drop of f_A minus that of f_B), the theorem caps $s \leq \alpha_A^P + \alpha_B^P + 2\hat{\eta}(\delta)$. Both experiments measure s against this cap: Section 5.2 where it decides the ordering, Section 5.1 where it does not.

Table 1: Model families and evaluation settings. Budgets use $[0, 1]$ pixel units; property sets are shared within each family.

	MNIST-FC	GTSRB-CNN	CIFAR-CNN
Architecture / models	3-layer FC / Plain+3	Small CNN / Plain+3	2 conv.+2 FC / Plain+6
Objective ($\varepsilon_{\text{train}}$)	PGD-AT ($\{.01, .02, .03\}$)	TRADES ($\{.01, .02, .03\}$)	PGD-AT ($\{.5, 1, 1.5, 2, 2.5, 3\}/255$)
Clean accuracy	92.6–93.2%	87.3–89.3%	58.1–58.8%
ID / structural OOD	MNIST / EMNIST-digits	GTSRB / BTSD	CIFAR-10 / —
Corruption OOD	PixMix	PixMix	CIFAR-10-C
Properties (ID / corr.)	100 / 100	215 / 420	100 / 100
Verification (ε; timeout)	$\{1, 2, 3, 4\}/255$; 60 s	$\{1, 2, 3, 4\}/255$; 60 s	$\{1, 2\}/255$; 120 s

3.3 From CRA to Out-of-Distribution Performance

The same coupling argument bounds the OOD value of any loss that depends only on the predicted label by ID CRA at radius δ and the defect $\hat{\eta}(\delta)$; it is the theoretical side of Section 5.3.

Definition 4 (Prediction-induced loss)**.** *A function $\ell : \mathbb{R}^K \times \mathcal{Y} \to [0, M]$ is* prediction-induced *if there exists $g : \mathcal{Y} \times \mathcal{Y} \to [0, M]$ with $g(y, y) = 0$ such that $\ell(\mathbf{u}, y) = g(\arg\max_j u_j,\ y)$ for every $\mathbf{u}, y$. The empirical loss of f_θ on a sample $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ is $L^\ell_{\mathcal{S}}(f_\theta) := N^{-1} \sum_i \ell(f_\theta(\mathbf{x}_i), y_i)$.*

The class of prediction-induced losses includes 0/1 error, top-k error, and any aggregation thereof (e.g., macro-F1 read as a weighted sum of per-class accuracies). Probabilistic metrics that depend on the full softmax (NLL, multiclass Brier) are *not* prediction-induced; they are addressed by the Lipschitz-augmented confidence-floor extension in Appendix A.7, which adds an $L_\ell L_f \delta$ slack term but recovers the same shape of bound. ECE is a binned statistic of the joint sample (not a pointwise loss $\mathbb{R}^K \times \mathcal{Y} \to [0, M]$) and is therefore outside the framework; we report it empirically in Section 5.3.

Theorem 2 (CRA-bounded loss shift)**.** *Under Assumptions 1–2, for every prediction-induced loss ℓ with bound M and every $\delta \geq 0$,*

$$L^\ell_{\mathcal{S}_Q}(f_\theta) \ \leq \ M \cdot \big[\, (1 - \mathrm{CRA}(f_\theta, \mathcal{S}_P, \delta)) + \hat{\eta}(\delta) \,\big]. \tag{11}$$

Corollary 2 (OOD accuracy lower bound)**.** *Specializing ℓ to the $0/1$ loss yields*

$$\mathrm{acc}(f_\theta, \mathcal{S}_Q) \ \geq \ \mathrm{CRA}(f_\theta, \mathcal{S}_P, \delta) \ - \ \hat{\eta}(\delta). \tag{12}$$

Corollary 3 (Pairwise OOD-loss rank bound)**.** *For two classifiers f_A, f_B, any prediction-induced loss ℓ, and every $\delta \geq 0$, the following bound holds:*

$$\begin{aligned} L^\ell_{\mathcal{S}_Q}(f_A) - L^\ell_{\mathcal{S}_Q}(f_B) \leq \big(L^\ell_{\mathcal{S}_P}(f_A) - L^\ell_{\mathcal{S}_P}(f_B)\big) \\ + M\Big[(1 - \mathrm{CRA}(f_A, \mathcal{S}_P, \delta)) + (1 - \mathrm{CRA}(f_B, \mathcal{S}_P, \delta)) + \hat{\eta}(\delta)\Big]. \end{aligned} \tag{13}$$

Between two clean-accuracy-matched classifiers, the OOD-loss disadvantage of f_A is thus bounded by its ID-loss disadvantage plus a slack that shrinks as either model's ID CRA grows. The proofs are a per-index dichotomy on Proposition 1 (Appendix A.6), machine-checked in Lean 4 (Appendix A.5).

Checking the Assumptions on a New Shift. The framework is a diagnostic that needs OOD inputs but no OOD verification. **(i)** Check that the label marginals agree, or add $\mathrm{TV}(P_Y, Q_Y)$ to the defect. **(ii)** Measure $\hat{\eta}(\delta)$: under the identity coupling when the shift is a per-input transformation, and otherwise from class-restricted nearest-neighbour distances in the certificate norm, which lower-bound the transport any label-preserving coupling must pay. **(iii)** Compare the budget (10) with the ID gap over δ. If the budget is met, the ordering is guaranteed; if it is not, the theorem still caps the erosion s, and the ordering must be measured.

4 Experimental Setup

Models. To study Corollary 1 empirically, we train a set of models on the same training data so that each family consists of models that share an architecture and are matched on in-distribution test accuracy but differ in in-distribution CRA. We obtain this spread by training one model with the plain cross-entropy objective and the others with adversarial training, PGD-AT (Madry et al., 2018) or TRADES (Zhang et al., 2019), at increasing $\varepsilon_{\text{train}}$, with accuracy-based early stopping to enforce the match. Table 1 summarizes the families; per-model configurations and clean accuracies are in

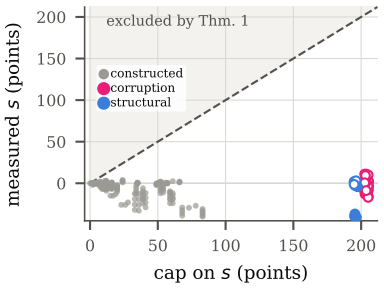


Figure 2: Measured erosion s against its cap in Theorem 1 (shaded: excluded). Grey: 206 constructed pair-cells; colour: 70 natural-shift pairs at cap 200, offset for visibility; filled: significant ID gap.

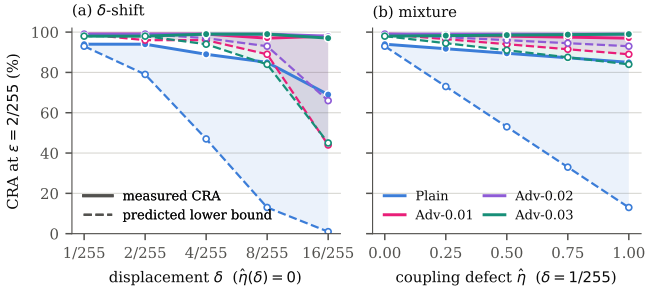


Figure 3: Measured CRA (solid) against the predicted lower bound (dashed) for the MNIST-FC models at $\varepsilon = 2/255$; the shaded band is the slack. (a) Every input moved by exactly δ, so $\hat{\eta}(\delta) = 0$ and the bound is $\mathrm{CRA}(\mathcal{S}_P, \varepsilon + \delta)$. (b) A fraction $\hat{\eta}$ of the inputs moved by $8/255$ and the rest by $1/255$, so the defect is $\hat{\eta}$ by construction.

Table 6 (Appendix C.2).

Test Distributions. We use the standard MNIST, GTSRB and CIFAR-10 test splits as in-distribution (ID) sets and consider two OOD shifts. (i) *Structural OOD* preserves the label space and input format but changes the data source: EMNIST-digits (Cohen et al., 2017), orientation-corrected to MNIST's layout, for MNIST, and a class-balanced subset of BTSD (Timofte et al., 2014), relabeled to the GTSRB taxonomy, for GTSRB. (ii) *Corruption OOD* simulates real-world image degradations (noise, blur, lighting) by applying PixMix (Hendrycks et al., 2022) to the ID test images (MNIST-PixMix, GTSRB-PixMix) and, for CIFAR-10, by pooling CIFAR-10-C (Hendrycks & Dietterich, 2019) (corruption types and severities were not retained). We retain only images that all models in a family classify correctly, so that CRA differences do not reflect differences in clean correctness. Figure 6 (Appendix D.1) shows examples.

Verification Protocol. To verify CRA on in-distribution and OOD data, we represent each query $(\mathbf{x}, k)$ in $\mathcal{S}_P$ or $\mathcal{S}_Q$ as a local ℓ_∞-robustness property in VNNLIB format (Bak et al., 2021), requiring $\arg\max_j f_j(\mathbf{x}') = k$ for all $\mathbf{x}' \in [0,1]^d$ with $\|\mathbf{x}' - \mathbf{x}\|_\infty \leq \varepsilon$. So that the CRA gap between two models is measured on the same inputs, each property set is held fixed across every model in a family; $\mathcal{S}_P$ contains 100 properties for MNIST (10 per class), 215 for GTSRB (5 per class) and 100 for CIFAR-10. Every property is checked with α, β-CROWN (Wang et al., 2021) at $\varepsilon \in \{1, 2, 3, 4\}/255$ with a 60-second per-instance timeout (CIFAR: $\{1, 2\}/255$, 120 s; Table 1); the MNIST-FC suite is additionally checked by two complementary verifiers, MARABOU (Wu et al., 2024) and NNV (Lopez et al., 2023). Timeouts count as uncertified, and per-instance verdicts support paired comparisons.

OOD Performance Metrics. In addition to verifier-based CRA, we evaluate each model on the full test set of every distribution with six performance metrics: accuracy and macro-F1 (predictive quality), ECE, Brier score and NLL (calibration), and the AUROC of the maximum softmax probability for separating OOD from ID inputs (definitions in Appendix C.3).

5 Experimental Results

5.1 Experiment 1: Rank Preservation under Natural Shift

We compare ID and OOD CRA rankings within each model family to determine whether a certified robustness advantage persists under distribution shift. For a pair (f, g) with higher ID CRA for f, let s be the CRA drop of f minus that of g, where each drop is measured from ID to OOD at a fixed certification radius. Then $\Delta_{\mathcal{S}_Q} = \Delta_{\mathcal{S}_P} - s$, so the ID ordering is preserved, allowing ties, if and only if $s \leq \Delta_{\mathcal{S}_P}$; Theorem 1 caps s at $\alpha_A^P + \alpha_B^P + 2\hat{\eta}(\delta)$. We compute these gaps from per-instance verification outcomes on shared property sets and test ID differences using an exact McNemar test ($p < 0.05$; Appendix D.2). Comparisons are made separately for each radius and OOD distribution; pairs with equal ID CRA are excluded because they have no ordering to preserve.

All 33 comparisons with a statistically significant ID CRA gap preserve their ordering under shift. These account for 33 of the 70 comparisons with distinct ID CRA, and the smallest significant gap is 6 percentage points. Among the 37 comparisons without a significant ID difference, 31 preserve their ordering. The 6 inversions occur in the CIFAR families and involve ID gaps of at most 5 points ($p \geq 0.42$); their OOD gaps are also non-significant ($p \geq 0.09$). Although absolute CRA varies

Table 2: CRA (%) at $\varepsilon \in \{1, 2\}/255$ for the MNIST-FC and GTSRB-CNN families on shared property sets under ID, structural OOD (EMNIST, BTSD), and corruption OOD (PixMix) conditions. Clean accuracy is evaluated on the full ID test set. **Bold** indicates the highest CRA in each column within a block. Per-cell standard errors are at most 5 percentage points (N=100) or 3.4 points (N=215). CIFAR results appear in Table 8.

Family	Verifier	Model	Clean acc.	In-distribution		Structural OOD		Corruption OOD	
				$^1/_{255}$	$^2/_{255}$	$^1/_{255}$	$^2/_{255}$	$^1/_{255}$	$^2/_{255}$
MNIST-FC	α, β-CROWN	Plain	93.2	92	89	54	52	63	58
		Adv-0.01	93.2	**95**	**93**	55	**55**	**73**	68
		Adv-0.02	92.6	94	**93**	**57**	54	70	**70**
		Adv-0.03	92.6	93	**93**	55	53	70	68
	Marabou	Plain	93.2	92	89	54	52	63	58
		Adv-0.01	93.2	**95**	**93**	55	**55**	**73**	68
		Adv-0.02	92.6	94	**93**	**57**	54	70	**70**
		Adv-0.03	92.6	93	**93**	55	53	70	68
	NNV	Plain	93.2	92	89	54	52	63	58
		Adv-0.01	93.2	**95**	**93**	55	**55**	**73**	68
		Adv-0.02	92.6	94	**93**	**57**	54	70	**70**
		Adv-0.03	92.6	93	**93**	55	53	70	68
GTSRB-CNN	α, β-CROWN	Plain	89.3	27	8	78	14	13	3
		TRADES-0.01	89.0	58	25	**98**	76	65	37
		TRADES-0.02	87.2	**63**	**32**	**98**	**90**	**67**	**40**
		TRADES-0.03	88.1	60	30	**98**	83	66	36

substantially across distributions, the undefended model has the lowest OOD CRA within each family (Tables 2 and 8; Appendix D.1).

Figure 2 places each slack against the cap. On every natural axis $\hat{\eta} = 1$, so Theorem 1 allows a shift to erode an ordering by 200 points; the measured slack has a median of -3 and a maximum of 11 points, and for every pair with a significant ID gap it stays below that gap. The full model rankings show similar agreement: Spearman's ρ between ID and OOD CRA in the seven-model CIFAR family is 0.96 $[0.56, 0.99]$ at $2/255$ and 0.84 $[0.33, 0.96]$ at $1/255$ (Appendix D.3). Across the 17-model MNIST-FC ablation, it is 0.77 on EMNIST and 0.86 on PixMix (Appendix E.1).

Conclusion. ID CRA is a useful criterion for comparing certified robustness under the natural shifts studied here: every statistically significant ID advantage is retained, and the observed reversals involve only non-significant differences. The agreement is measured, not guaranteed: the cap on s exceeds every ID gap here. What the theorem supplies is the scale. Natural shifts could erode an ordering by 200 points, yet erode none by more than 11, a small fraction of the worst case they admit. Section 5.2 tests the cap where it is small enough to decide the ordering.

5.2 Experiment 2: Where the Guarantee Applies

We first evaluate the transfer bound under controlled shifts with known defect, providing a direct test of its quantitative prediction. The defect estimates for the natural-shift axes, including their dependence on the evaluated displacement range, are reported in Appendix D.5.

Figure 3 compares two constructions on 100 shared MNIST inputs. First, displacements bounded by δ in ℓ_∞ yield $\hat{\eta}(\delta) = 0$, giving the lower bound $\mathrm{CRA}(f, \mathcal{S}_Q, \varepsilon) \geq \mathrm{CRA}(f, \mathcal{S}_P, \varepsilon + \delta)$. Second, a mixture of small and large displacements varies the defect from zero to one (Appendix D.4). Together, the constructions assess tightness as displacement and defect increase.

At $\varepsilon = 2/255$, measured CRA meets the zero-defect bound in all 20 cells, with 0–3 percentage points of slack for robust models when $\delta \leq 2/255$. Plain has the lowest CRA at every δ and is significantly below each robust model for $\delta \geq 4/255$ under paired McNemar tests ($p \leq 0.006$); differences among the robust models are not significant. Across the full sweep of 320 cells spanning two architecture families, three radii, and both constructions, no measured CRA falls below the bound. Slack increases with the defect, reflecting the bound's worst-case penalty for mass transported beyond δ.

Conclusion. The bound is valid and, per model, tight; slack grows predictably with the defect. Per pair (Figure 2, grey), s never exceeds the cap in 206 pair-cells, and the cap decides the ordering before any OOD verification in 9, all of which hold.

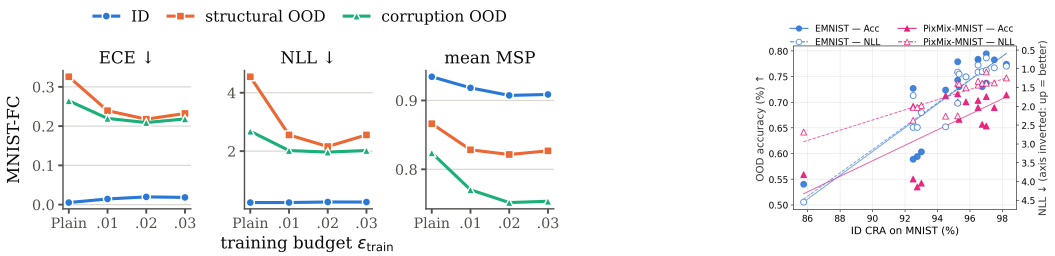


Figure 4: ID CRA as a proxy for OOD performance. **(a)** ECE, NLL and mean MSP for the MNIST-FC family against training budget, one line per suite (lower is better for ECE and NLL; mean MSP should have ID above OOD): robust training improves probability quality under shift, saturating after the first budget. GTSRB-CNN is in Appendix E.2. **(b)** ID CRA on MNIST against OOD accuracy and NLL (inverted) across the 17-model ablation.

5.3 Experiment 3: ID CRA as a Model-Selection Proxy

We ask whether a model selected by higher in-distribution CRA also performs better on shifted data. We therefore compare ID CRA with six OOD metrics across the 17-model MNIST-FC ablation. Spearman correlations summarize the direction and strength of these associations; with $n = 17$, they are descriptive trends, not significance tests.

For the natural-shift axes, we therefore evaluate the relationship between ID CRA and OOD behavior empirically. The strongest associations are with predictive loss. ID CRA correlates with lower OOD NLL on EMNIST ($\rho = -0.90$) and PixMix ($\rho = -0.89$), and with higher OOD accuracy ($\rho = +0.78$ and $+0.41$, respectively). It also tracks max-softmax OOD-detection AUROC, but less consistently: $\rho = +0.44$ on EMNIST and $+0.92$ on PixMix. Thus, ID CRA can help select a model whose existing confidence-based detector performs well; it is not itself an input-level OOD score. The metric breakdown in Figure 4a and Appendix E.2 shows that adversarial training improves NLL across all OOD suites, whereas calibration does not improve uniformly on BTSD.
We finally ask what a deployer gives up by choosing the model with the highest ID CRA. In four of six settings this choice is also best OOD, and in a fifth it ties for best. The largest observed miss is on CIFAR-10-C at $\varepsilon = 2/255$: the ID-CRA choice certifies 80% of the properties, whereas the OOD-best model certifies 87%—seven additional certified examples per 100 properties. This is the observed selection cost in our experiments, not a universal upper bound; a random choice loses 2–20 percentage points on these settings. Corollary 3 gives the corresponding theoretical ordering result.
Conclusion. ID CRA is informative for model selection under shift, especially for OOD loss and accuracy, but it should be treated as a proxy with a measurable, shift-dependent selection cost.

6 Limitations and Conclusion

Limitations. Our empirical scope is the verification-tractable regime (MNIST-FC, GTSRB-CNN, CIFAR-CNN). The shift defect is exact for PixMix and a lower bound for EMNIST and BTSD (Appendix D.5); it equals one at the certified radii on every natural axis, so the rank agreement there is measured, not guaranteed. The bounds are worst case; a distributional analysis may explain why natural shifts stay far inside them. Supported here: verification-tractable ReLU networks under label-preserving covariate shift, across architecture, objective, normalization and verifier. Plausible but untested: larger architectures under the same kinds of shift. Expected to fail: semantic or label shift, open-set inputs, and shifts severe enough that CRA collapses for every model.
Conclusion. A deployer with one in-distribution certificate can now say what it licenses. If the shift keeps most inputs within the certified radius, the check of Section 3 establishes it and Theorem 1 guarantees the ordering; on the natural shifts we study the check reports that it does not, the theorem still caps how far the ordering can erode, and the measured erosion stays far inside that cap: the CRA ranking shows strong but imperfect agreement, ID CRA tracks OOD loss, calibration and detection, and selecting by it costs at most a few points of OOD CRA. The certificate does not travel; the ordering it induces largely does.

AI USE STATEMENT

We used generative AI coding assistants (Claude Code) as tools under the authors' direction, mainly to help write and debug experiment and analysis scripts, and to edit and proofread the manuscript's text and LaTeX. The research questions, study design, theoretical results, interpretation of findings and conclusions are the authors' own. Every reported number comes from executed code and logged verifier outputs. The authors reviewed all AI-assisted code and text and take full responsibility for the content of this paper.

ETHICS STATEMENT

This work uses only public benchmark image datasets (MNIST, EMNIST, GTSRB, BTSD, and their PixMix-corrupted variants) and standard verification tooling; it involves no human subjects, crowdsourcing, or sensitive data. The study asks whether certified robustness guarantees, which are used to deploy ML systems in safety-critical settings, transfer under distribution shift. The intended positive impact is more reliable deployment of certified defenses without re-verifying every shifted distribution. We do not see a direct misuse path: certified-robustness analysis is a defensive primitive, and our findings make existing defenses easier to evaluate rather than making stronger attacks easier to mount. The Limitations paragraph (Section 6) cautions against extrapolating the empirical findings beyond the verification-tractable regime studied here.

REPRODUCIBILITY STATEMENT

An anonymized code repository, linked in the abstract, contains the training scripts, VNNLIB property generation, and verification driver scripts for α, β-CROWN, Marabou, and NNV. Model architectures are specified in Appendix C.1; training procedures and the clean-accuracy-matched model configurations are given in Appendix C.2 (Table 6), and compute resources in Appendix C.2. The shared property sets and verifier configurations are described in Section 5 and Appendix F.1. All theoretical results state their assumptions explicitly and are proved in Appendix A; the Lean 4 / mathlib formalization is described in Appendix A.5, and its sources are provided as supplementary material. All datasets used are public.

REFERENCES

Brendon G Anderson and Somayeh Sojoudi. Data-driven certification of neural networks with random input noise. *IEEE transactions on control of network systems*, 10(1):249–260, 2022.

Maximilian Augustin, Alexander Meinke, and Matthias Hein. Adversarial robustness on in- and out-distribution improves explainability. In *European Conference on Computer Vision*, pp. 228–245. Springer, 2020.

Stanley Bak, Changliu Liu, and Taylor Johnson. The second international verification of neural networks competition (vnn-comp 2021): Summary and results. *arXiv preprint arXiv:2109.00498*, 2021.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. EMNIST: Extending MNIST to handwritten letters. *Proceedings of the International Joint Conference on Neural Networks*, 2017.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pp. 1310–1320. PMLR, 2019.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning (ICML)*, 2020.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets*

and Benchmarks Track (Round 2), 2021. URL https://openreview.net/forum?id=SSKZPJCt7B.

Alex J DeGrave, Joseph D Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021.

Joy Dhar, Manish Kumar Pandey, Behzad Bozorgtabar, Nayyar Zaidi, Wenyu Zhang, Wei-Hong Li, Tingting Mu, Dwarikanath Mahapatra, Mahsa Baktashmotlagh, Trung Le, et al. Does a hybrid space-aware randomized defense improve empirical and certified adversarial robustness? In *Forty-third International Conference on Machine Learning*, 2026.

Angelos Filos, Panagiotis Tigkas, Rowan McAllister, Nicholas Rhinehart, Sergey Levine, and Yarin Gal. Can autonomous vehicles identify, recover from, and adapt to distribution shifts? In *International Conference on Machine Learning*, pp. 3145–3153. PMLR, 2020.

Timon Gehr, Matthew Mirman, Dana Drachsler-Cohen, Petar Tsankov, Swarat Chaudhuri, and Martin Vechev. Ai2: Safety and robustness certification of neural networks with abstract interpretation. In *2018 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2018.

Sven Gowal, Krishnamurthy Dj Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. Scalable verified training for provably robust image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4842–4851, 2019.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

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

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8320–8329. IEEE, 2021a.

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*, 2021b.

Dan Hendrycks, Andy Zou, Mantas Mazeika, Leonard Tang, Bo Li, Dawn Song, and Jacob Steinhardt. Pixmix: Dreamlike pictures comprehensively improve safety measures. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16762–16771. IEEE, 2022.

Taylor T. Johnson. Is neural network verification useful and what is next? In *2025 61st Allerton Conference on Communication, Control, and Computing Proceedings*. Allerton Conference on Communication, Control, and Computing, 2025. URL https://hdl.handle.net/2142/130315.

Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks. In *International conference on computer aided verification*, pp. 97–117. Springer, 2017.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664, 2021.

Aounon Kumar, Alexander Levine, Tom Goldstein, and Soheil Feizi. Provable robustness against wasserstein distribution shifts via input randomization. In *The Eleventh International Conference on Learning Representations*, 2023.

Linyi Li, Tao Xie, and Bo Li. Sok: Certified robustness for deep neural networks. In *2023 IEEE symposium on security and privacy (SP)*, pp. 1289–1310. IEEE, 2023.

Diego Manzanas Lopez, Sung Woo Choi, Hoang-Dung Tran, and Taylor T Johnson. Nnv 2.0: The neural network verification tool. In *International Conference on Computer Aided Verification*, pp. 397–412. Springer, 2023.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *Conference on Learning Theory*, 2009.

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

Matthew Mirman, Timon Gehr, and Martin Vechev. Differentiable abstract interpretation for provably robust neural networks. In *International Conference on Machine Learning*, 2018.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Skz_WfbCZ.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning*, 2019.

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifiable distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Hk6kPgZA-.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks*, 32:323–332, 2012.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in neural information processing systems*, 33:18583–18599, 2020.

Radu Timofte, Karel Zimmermann, and Luc Van Gool. Multi-view traffic sign detection, recognition, and 3d localisation. *Machine vision and applications*, 25(3):633–647, 2014.

Vincent Tjeng, Kai Y. Xiao, and Russ Tedrake. Evaluating robustness of neural networks with mixed integer programming. In *International Conference on Learning Representations (ICLR)*, 2019.

Hoang-Dung Tran, Xiaodong Yang, Diego Manzanas Lopez, Patrick Musau, Luan Viet Nguyen, Weiming Xiang, Stanley Bak, and Taylor T Johnson. Nnv: the neural network verification tool for deep neural networks and learning-enabled cyber-physical systems. In *International conference on computer aided verification*, pp. 3–17. Springer, 2020.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Shiqi Wang, Huan Zhang, Kaidi Xu, Xue Lin, Suman Jana, Cho-Jui Hsieh, and J Zico Kolter. Beta-crown: Efficient bound propagation with per-neuron split constraints for neural network robustness verification. *Advances in neural information processing systems*, 34:29909–29921, 2021.

Haoze Wu, Omri Isac, Aleksandar Zeljić, Teruhiro Tagomori, Matthew Daggitt, Wen Kokke, Idan Refaeli, Guy Amir, Kyle Julian, Shahaf Bassan, et al. Marabou 2.0: a versatile formal analyzer of neural networks. In *International Conference on Computer Aided Verification*, pp. 249–264. Springer, 2024.

Kaidi Xu, Huan Zhang, Shiqi Wang, Yihan Wang, Suman Jana, Xue Lin, and Cho-Jui Hsieh. Fast and complete: Enabling complete neural network verification with rapid and massively parallel incomplete verifiers. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=nVZtXBI6LNn.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pp. 7472–7482. PMLR, 2019.

Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel. Efficient neural network robustness certification with general activation functions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Huan Zhang, Shiqi Wang, Kaidi Xu, Linyi Li, Bo Li, Suman Jana, Cho-Jui Hsieh, and J. Zico Kolter. General cutting planes for bound-propagation-based neural network verification. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

A Theoretical Analysis: Proofs and Formalization

This appendix collects the proofs of the main-body results and the Lean 4 / mathlib formalization that machine-checks them. Appendix A.1 proves the CRA-shift lemma; Appendix A.3 proves Theorem 1 and Corollary 1; Appendix A.4 gives the two-sided form; Appendix A.6 proves Theorem 2 and its corollaries; Appendix A.7 states the confidence-floor extension (Theorem 3); Appendix A.8 characterizes when the bound is informative; Appendix A.5 maps every claim onto a closed Lean theorem. The remaining appendices follow the order of the main text: experimental details (Appendix C), the detailed analysis behind Experiments 1 and 2 (Appendix D), the ablations and model-selection analysis behind Experiment 3 (Appendix E), and the checks of Section 5.3 (Appendix F). Figure 1 (Section 1) illustrates the certified ball, the in-distribution CRA gap, and the rank-preservation prediction under shift.

A.1 CRA Shift Bound under a Coupled Sample

We first state the per-classifier shift bound that underlies all subsequent results. For a classifier f_θ and a fixed coupled pair of samples $\mathcal{S}_P$ and $\mathcal{S}_Q$ as in Section 3.2, the empirical shift defect $\hat{\eta}(\delta)$ is the index fraction $\hat{\eta}(\delta) = N^{-1}\big|\{i : \|\mathbf{x}'_i - \mathbf{x}_i\|_\infty > \delta\}\big|$.

Lemma 2 is stated in Section 3.2. Write $R_i(\varepsilon)$ for the event that f_θ is locally robust at $(\mathbf{x}_i, y_i)$ at radius ε, and $R'_i(\varepsilon)$ for the same at $(\mathbf{x}'_i, y_i)$, so that $\mathrm{CRA}(f_\theta, \mathcal{S}_P, \varepsilon) = N^{-1}|\{i : R_i(\varepsilon)\}|$ and $\mathrm{CRA}(f_\theta, \mathcal{S}_Q, \varepsilon) = N^{-1}|\{i : R'_i(\varepsilon)\}|$.

Proof of Lemma 2. (i) Lower bound (5). Consider any index $i \in \{1, \dots, N\}$ with $\|\mathbf{x}'_i - \mathbf{x}_i\|_\infty \le \delta$ and $R_i(\varepsilon + \delta)$. Every $\mathbf{z} \in [0,1]^d$ with $\|\mathbf{z} - \mathbf{x}'_i\|_\infty \le \varepsilon$ satisfies $\|\mathbf{z} - \mathbf{x}_i\|_\infty \le \varepsilon + \delta$ by the triangle inequality, so $\hat{y}(\mathbf{z}) = y_i$; that is, $R'_i(\varepsilon)$ holds. Hence

$$\big\{i : R_i(\varepsilon + \delta)\big\} \;\subseteq\; \big\{i : R'_i(\varepsilon)\big\} \;\cup\; \big\{i : \|\mathbf{x}'_i - \mathbf{x}_i\|_\infty > \delta\big\}.$$

Taking cardinalities and dividing by N yields $\mathrm{CRA}(f_\theta, \mathcal{S}_P, \varepsilon + \delta) \le \mathrm{CRA}(f_\theta, \mathcal{S}_Q, \varepsilon) + \hat{\eta}(\delta)$, which rearranges to (5).

(ii) Upper bound (6). Part (i) uses only that each pair shares its label and the displacement $\|\mathbf{x}'_i - \mathbf{x}_i\|_\infty$, which is symmetric in the two samples. Applying (i) with the roles of $\mathcal{S}_P$ and $\mathcal{S}_Q$ exchanged, with

the same $\hat{\eta}(\delta)$, gives $\mathrm{CRA}(f_\theta, \mathcal{S}_Q, \varepsilon + \delta) \le \mathrm{CRA}(f_\theta, \mathcal{S}_P, \varepsilon) + \hat{\eta}(\delta)$, which is (6). In Lean, (i) and (ii) are `empiricalCRA_shift_count` and `empiricalCRA_shift_count_upper`. □

Remark 1 (Pointwise form)**.** *Writing $d_i := \|\mathbf{x}'_i - \mathbf{x}_i\|_\infty$, the same argument applied to each index with its own displacement gives*

$$\mathrm{CRA}(f_\theta, \mathcal{S}_Q, \varepsilon) \;\ge\; \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\big[\rho_{f_\theta}(\mathbf{x}_i, y_i) - d_i > \varepsilon\big], \tag{14}$$

since $\rho_{f_\theta}(\mathbf{x}_i, y_i) - d_i > \varepsilon$ implies $\rho_{f_\theta}(\mathbf{x}'_i, y_i) > \varepsilon$ by Proposition 1, and hence local robustness at $(\mathbf{x}'_i, y_i)$ at radius ε.

A.2 Existence of Label-Preserving Couplings, and Label Shift

Lemma 1 is stated in Section 3.2.

Proof of Lemma 1. (i) If $P_Y = Q_Y$, draw Y from the common label law and then, independently, $\mathbf{x} \sim P(\cdot \mid Y)$ and $\mathbf{x}' \sim Q(\cdot \mid Y)$; the result has marginals P and Q and $Y = Y'$ surely. Conversely, if $Y = Y'$ π-a.s., the laws of Y and Y' coincide, and these are P_Y and Q_Y. (ii) Take a maximal coupling of (P_Y, Q_Y), under which $Y \ne Y'$ with probability $\mathrm{TV}(P_Y, Q_Y)$, and extend it as in (i) by sampling $\mathbf{x}$ and $\mathbf{x}'$ from the respective conditionals. The proofs of Lemma 2, Theorem 1, and Theorem 2 use, for each index, only that the displacement is at most δ and that the two labels agree; every index violating either condition is absorbed into the defect term, which proves the stated replacement. □

A.3 Proof of the Rank-Preservation Theorem

Proof of Theorem 1. Definition 1 allows any real radius (for $r < 0$ the ball is empty, so $\mathrm{CRA}(\cdot, \cdot, r) = 1$), and the proof of Lemma 2 holds verbatim for every $\varepsilon \in \mathbb{R}$. Apply Lemma 2 to f_A in form (5), and to f_B in form (6) with the lemma's ε replaced by $\varepsilon - \delta$:

$$\begin{aligned} \mathrm{CRA}(f_A, \mathcal{S}_Q, \varepsilon) \;&\ge\; \mathrm{CRA}(f_A, \mathcal{S}_P, \varepsilon + \delta) - \hat{\eta}(\delta), \\ \mathrm{CRA}(f_B, \mathcal{S}_Q, \varepsilon) \;&\le\; \mathrm{CRA}(f_B, \mathcal{S}_P, \varepsilon - \delta) + \hat{\eta}(\delta). \end{aligned}$$

Subtracting,

$$\Delta_{\mathcal{S}_Q}(\varepsilon) \;\ge\; \mathrm{CRA}(f_A, \mathcal{S}_P, \varepsilon + \delta) - \mathrm{CRA}(f_B, \mathcal{S}_P, \varepsilon - \delta) - 2\hat{\eta}(\delta).$$

Substituting the modulus-term definitions $\alpha_A^P = \mathrm{CRA}(f_A, \mathcal{S}_P, \varepsilon) - \mathrm{CRA}(f_A, \mathcal{S}_P, \varepsilon + \delta)$ and $\alpha_B^P = \mathrm{CRA}(f_B, \mathcal{S}_P, \varepsilon - \delta) - \mathrm{CRA}(f_B, \mathcal{S}_P, \varepsilon)$ gives (7). Both moduli are non-negative because CRA is non-increasing in ε (Definition 2). □

Proof of Corollary 1. Direct from (7): if $\Delta_{\mathcal{S}_P}(\varepsilon) > \alpha_A^P + \alpha_B^P + 2\hat{\eta}(\delta)$ then $\Delta_{\mathcal{S}_Q}(\varepsilon) > 0$. □

A.4 Two-Sided Form of the Rank-Preservation Bound

Corollary 4 (Two-sided gap-stability bound)**.** *Under the assumptions of Theorem 1,*

$$\Delta_{\mathcal{S}_P}(\varepsilon) - \alpha_A^P - \alpha_B^P - 2\hat{\eta}(\delta) \;\le\; \Delta_{\mathcal{S}_Q}(\varepsilon) \;\le\; \Delta_{\mathcal{S}_P}(\varepsilon) + \alpha_A^Q + \alpha_B^Q + 2\hat{\eta}(\delta), \tag{15}$$

where the OOD-side moduli are $\alpha_A^Q := \mathrm{CRA}(f_A, \mathcal{S}_Q, \varepsilon) - \mathrm{CRA}(f_A, \mathcal{S}_Q, \varepsilon + \delta)$ and $\alpha_B^Q := \mathrm{CRA}(f_B, \mathcal{S}_Q, \varepsilon - \delta) - \mathrm{CRA}(f_B, \mathcal{S}_Q, \varepsilon)$.

Proof. The lower bound is Theorem 1. The upper bound is obtained by applying Theorem 1 with $\mathcal{S}_P$ and $\mathcal{S}_Q$ swapped (the empirical shift defect is symmetric in the two samples) and rewriting the resulting inequality in terms of α_A^Q, α_B^Q. □

The corollary captures four conceptual relationships between $\Delta_{\mathcal{S}_P}$ and $\Delta_{\mathcal{S}_Q}$ that are testable on the empirical results in Section 5: (i) $\Delta_{\mathcal{S}_Q}$ lies in a window centered on $\Delta_{\mathcal{S}_P}$ with half-width controlled by the modulus terms and $2\hat{\eta}(\delta)$; (ii) rank preservation is guaranteed whenever $|\Delta_{\mathcal{S}_P}| > \alpha_A^P + \alpha_B^P + 2\hat{\eta}(\delta)$, with the Q-side moduli not entering this sufficient condition (this is what makes the prediction *a priori*); (iii) as $(\delta, \hat{\eta}(\delta)) \to 0$ the modulus terms also vanish and the gaps coincide; (iv) for fixed $(\delta, \hat{\eta}(\delta))$, $\Delta_{\mathcal{S}_Q}$ rises with $\Delta_{\mathcal{S}_P}$ at slope one with offset bounded by the budget, the certified-accuracy analogue of the empirical "accuracy on the line" regularity (Taori et al., 2020; Miller et al., 2021). The GTSRB-to-BTSD axis (Section 5.1), where every model has higher absolute CRA OOD than ID, is consistent with the two-sided bound, which permits such increases. This does not establish the sufficient condition for rank preservation. On BTSD, as on every natural axis tested, the measured defect is one over the evaluated displacement range (Appendix D.5), so the budget exceeds every ID gap. The observed BTSD ordering is empirical; the framework does not guarantee it.

Couplings and distance diagnostics. The shift defect $\hat{\eta}(\delta)$ is defined relative to a coupling between $\mathcal{S}_P$ and $\mathcal{S}_Q$. For PixMix, pairing each ID image with its own corrupted variant gives a valid coupling and an exactly measurable defect. For EMNIST, BTSD, and the pooled CIFAR-10-C subset, we instead measure each shifted image's distance to its nearest same-class source image in the certificate norm. This assignment need not preserve the source marginal, so it is not in general a coupling of the two empirical distributions. The distances lower-bound the displacement under any label-preserving coupling on these samples; their exceedance fraction therefore lower-bounds $\hat{\eta}(\delta)$ (Appendix D.5). A lower bound of one establishes that the defect is one, as observed at the evaluated radii. A smaller lower bound cannot certify that the budget is met: that requires a valid coupling and an upper bound on its defect. The controlled shifts in Section 5.2 supply explicit pairings and known defects.

A.5 Lean 4 / mathlib Formalization

Every formal claim above (Proposition 1, Lemma 2, Theorems 1, 2, and 3, and Corollaries 1, 4, 2, and 3) is machine-checked in Lean 4 against mathlib at a pinned commit; the Lean source (`CRAGeneralization.lean`, `CRAGenPopulation.lean`, and `AxiomCheck.lean`) is provided as supplementary material. Each Lean theorem builds without `sorry` and depends only on the standard mathlib axioms `propext`, `Classical.choice`, and `Quot.sound`. The Lean development states local robustness on balls in the whole space; the paper's balls, restricted to $[0,1]^d$, are recovered by applying it to $\hat{y} \circ \mathrm{clip}$, where clip clamps each coordinate to $[0,1]$: for $\mathbf{x} \in [0,1]^d$ clamping does not increase any coordinate's distance to $\mathbf{x}$, so $\hat{y} \circ \mathrm{clip}$ is robust on a full ball exactly when $\hat{y}$ is robust on the corresponding clipped ball. Proposition 1 is checked for inputs whose radius sets are nonempty and bounded, i.e. both points correctly classified and $\hat{y}$ not constant on $[0,1]^d$. Table 3 records the paper-to-Lean correspondence; Corollary 1 is an immediate consequence of Theorem 1 and is covered by the same Lean theorem `empirical_rank_preservation_count`. The population/measure-theoretic versions of the loss-shift results, formulated against an arbitrary coupling of probability measures and the Bochner integral, are also machine-checked (`population_loss_id_bound`, `population_loss_shift_bound`, `population_loss_rank_bound`); they require additional measurability hypotheses on the local-robustness predicate but no new axioms beyond those listed above.

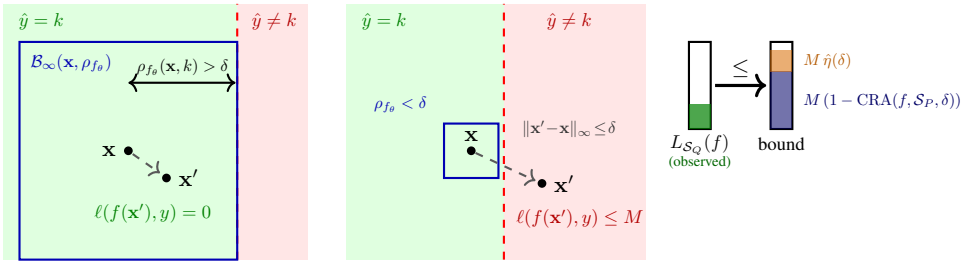
The Lean statements work in count form (numbers of indices satisfying each predicate in a finite-sample-of-size-N index family), which differs from the paper's CRA form only by an overall factor of $1/N$; the equivalence is mechanical and is the reason the modulus and shift-defect terms appear with identical coefficients in both representations.

A.6 Proof of the CRA-Bounded Loss Shift

Theorem 2 and Corollaries 2–3 extend the CRA-shift argument from indicator counts to general prediction-induced losses. The proofs are a per-index dichotomy that reuses Proposition 1 (or directly the geometric core `locallyRobust_shift`); the population-level statements follow by passing to the limit. Figure 5 illustrates the dichotomy at the per-input level and the resulting bound at the population level.

Table 3: Correspondence between the paper-level claims and the Lean theorems that machine-check them. The Lean theorems live in `CRAGeneralization.lean`; `AxiomCheck.lean` re-prints the axioms each theorem depends on.

Paper-level claim	**Lean theorem**
Proposition 1 (ρ_{f_θ} is 1-Lipschitz)	`certifiedRadius_lipschitz`
Per-input shift implication (geometric core of Proposition 1)	`locallyRobust_shift`
Pointwise indicator form of Lemma 2	`cra_indicator_shift`
One-sided ρ_{f_θ}-Lipschitz bound (used inside the symmetric argument)	`certifiedRadius_lipschitz_one_sided`
Set-level lifting of the per-input shift	`robustRadii_shift_subset_strong`
Symmetry of the empirical shift defect	`numShiftViolations_symm`
Lemma 2 (lower bound, (5))	`empiricalCRA_shift_count`
Lemma 2 (upper bound, (6))	`empiricalCRA_shift_count_upper`
Theorem 1 (pairwise rank preservation)	`empirical_rank_preservation_count`
Corollary 4 (upper-bound direction)	`empirical_rank_upper_bound`
Corollary 4 (two-sided bracket as a conjunction)	`empirical_gap_two_sided`
Helper: $\lvert\{i : p\}\rvert + \lvert\{i : \neg p\}\rvert = n$	`filter_card_add_neg_card`, `filter_neg_card_real`
Helper: numLR $\le n$ and $\hat{\eta}(\delta)$=0 at $\mathbf{x}'$=$\mathbf{x}$	`numLocallyRobust_le_n`, `numShiftViolations_self_zero`
Theorem 2 (CRA-bounded loss shift)	`empirical_loss_shift_bound`
Corollary 2 (OOD accuracy lower bound)	`empirical_accuracy_lower_bound`
ID-only specialization (helper for the rank corollary)	`empirical_loss_id_bound`
Corollary 3 (pairwise OOD-loss rank bound)	`empirical_loss_rank_bound`
Theorem 3 (Lipschitz-augmented confidence-floor; the Lean version takes the shifted-point bound directly as a hypothesis, the Lipschitz derivation feeds it)	`empirical_loss_shift_bound_floor`
Population/measure-theoretic versions (in `CRAGenPopulation.lean`)	
Population coupling structure (joint with prescribed marginals)	`Coupling`
Population expected loss $\mathbb{E}_Q[\ell(f(\mathbf{x}), y)]$	`populationLoss`
Bounded loss $\Rightarrow$ integrable on a finite measure (helper)	`integrable_loss_of_meas`
Population analogue of Theorem 2 (ID specialization)	`population_loss_id_bound`
Population analogue of Theorem 2 (full coupling form)	`population_loss_shift_bound`
Population analogue of Corollary 3 (pairwise rank)	`population_loss_rank_bound`



(a) Certified: $\mathbf{x}'$ stays inside the ball

(b) Uncertified: $\mathbf{x}'$ may cross the boundary

(c) $L_{\mathcal{S}_Q}(f) \le M\,((1-\mathrm{CRA}) + \hat{\eta})$

Figure 5: Conceptual illustration of the CRA-bounded loss-shift theorem (Theorem 2). **(a)** A point $\mathbf{x}$ with certified radius $\rho_{f_\theta}(\mathbf{x}, k) > \delta$ has the matched OOD point $\mathbf{x}'$ inside its certified ℓ_∞-ball, so the prediction is preserved and the prediction-induced loss vanishes by $g(y, y) = 0$. **(b)** A point with $\rho_{f_\theta}(\mathbf{x}, k) < \delta$ has a ball that may not contain $\mathbf{x}'$; the prediction can flip across the decision boundary, contributing at most M to the loss. **(c)** At the population level, the OOD expected loss $L_{\mathcal{S}_Q}(f)$ is bounded above by $M \cdot ((1 - \mathrm{CRA}(f, \mathcal{S}_P, \delta)) + \hat{\eta}(\delta))$, which decomposes into the uncertified-mass contribution $M\,(1-\mathrm{CRA}(f, \mathcal{S}_P, \delta))$ and the shift-defect contribution $M\,\hat{\eta}(\delta)$. Both components are observable from in-distribution evaluation: CRA from the verifier output at radius δ, $\hat{\eta}$ from the coupling. The bound's predictive content is best when CRA is large at a δ at which $\hat{\eta}$ is small.

Proof of Theorem 2. For each index $i \in \{1, \ldots, N\}$, let $E_i := \{R_i(\delta)\} \cap \{\|\mathbf{x}'_i - \mathbf{x}_i\|_\infty \le \delta\}$, with R_i as in Appendix A.1. On E_i, local robustness at radius δ gives $\hat{y}(\mathbf{x}'_i) = y_i$, hence $\ell(f_\theta(\mathbf{x}'_i), y_i) = g(y_i, y_i) = 0$ by the diagonal-zero hypothesis on g. On E_i^c, $\ell(f_\theta(\mathbf{x}'_i), y_i) \le M$. Summing,

$$N\,L^{\ell}_{\mathcal{S}_Q}(f_\theta) \;=\; \sum_{i \in E_i^c} \ell(f_\theta(\mathbf{x}'_i), y_i) \;\le\; M \cdot |E_i^c| \;\le\; M \cdot \big[\, N\,(1 - \mathrm{CRA}(f_\theta, \mathcal{S}_P, \delta)) + N\,\hat{\eta}(\delta)\,\big],$$

where the last inequality is the union bound on $E_i^c \subseteq \{\neg R_i(\delta)\} \cup \{\|\mathbf{x}'_i - \mathbf{x}_i\|_\infty > \delta\}$. Dividing by N yields (11). □

Proof of Corollary 2. Apply Theorem 2 with $\ell = 0/1$ loss ($M = 1$) and use $L^{0/1}_{\mathcal{S}_Q}(f_\theta) = 1 - \mathrm{acc}(f_\theta, \mathcal{S}_Q)$. Rearranging gives $\mathrm{acc}(f_\theta, \mathcal{S}_Q) \ge \mathrm{CRA}(f_\theta, \mathcal{S}_P, \delta) - \hat{\eta}(\delta)$. □

Proof of Corollary 3. Decompose the OOD-loss gap as

$$L^{\ell}_{\mathcal{S}_Q}(f_A) - L^{\ell}_{\mathcal{S}_Q}(f_B) = \big[L^{\ell}_{\mathcal{S}_Q}(f_A) - L^{\ell}_{\mathcal{S}_P}(f_A)\big] + \big[L^{\ell}_{\mathcal{S}_P}(f_A) - L^{\ell}_{\mathcal{S}_P}(f_B)\big] + \big[L^{\ell}_{\mathcal{S}_P}(f_B) - L^{\ell}_{\mathcal{S}_Q}(f_B)\big].$$

The first bracket is bounded above by $L^{\ell}_{\mathcal{S}_Q}(f_A)$ (since $L^{\ell}_{\mathcal{S}_P}(f_A) \geq 0$), which by Theorem 2 applied to f_A is at most $M\left[(1 - \mathrm{CRA}(f_A, \mathcal{S}_P, \delta)) + \hat{\eta}(\delta)\right]$. The third bracket is bounded above by $L^{\ell}_{\mathcal{S}_P}(f_B)$ (since $L^{\ell}_{\mathcal{S}_Q}(f_B) \geq 0$); applying Theorem 2 to f_B with $\mathcal{S}_P, \mathcal{S}_Q$ identified (the ID-only specialization, where the shift defect vanishes) gives the bound $M\left[1 - \mathrm{CRA}(f_B, \mathcal{S}_P, \delta)\right]$. The middle bracket is the ID-loss difference. Summing the three bounds yields (13). □

A.7 Confidence-Floor Extension to General Bounded Losses

Scope of the formalization. The Lean theorem in this subsection (`empirical_loss_shift_bound_floor`) is machine-checked at the level of the shifted-point hypothesis: it takes the bracketed inequality $\ell(f_\theta(\mathbf{x}'_i), y_i) \leq \varepsilon^{\mathrm{ID}}_{\mathrm{floor}} + L_\ell L_f \delta$ as a direct assumption. The Lipschitz derivation that supplies that hypothesis (Assumptions 4–5 together with Assumption 3) is paper-side reasoning, not Lean-checked.

Theorem 2 requires the loss to be prediction-induced, i.e., depend only on $\arg\max_j f_j(\mathbf{x})$ and the true label. This covers classification accuracy, top-k accuracy, and macro-F1, but not metrics that depend on the full softmax distribution (NLL, multiclass Brier). The following extension handles bounded losses $\ell : \mathbb{R}^K \times \mathcal{Y} \to [0, M]$ at the cost of two additional assumptions: a confidence floor on certified *source* points and Lipschitz continuity of the loss and the network. The extension is machine-checked in Lean as `empirical_loss_shift_bound_floor` (with the conclusion taken on the GOOD set; the Lipschitz derivation below feeds the right shifted-point bound into that hypothesis).

Why a naive source-point floor is not enough. A natural first attempt is to assume $\ell(f_\theta(\mathbf{x}), y) \leq \varepsilon_{\mathrm{floor}}$ on certified *source* points and conclude $L^{\ell}_{\mathcal{S}_Q} \leq \varepsilon_{\mathrm{floor}}\,\mathrm{CRA} + M\left((1-\mathrm{CRA}) + \hat{\eta}\right)$ by the same per-input dichotomy as Theorem 2. This conclusion is *false*: local robustness preserves the predicted class under shift, but it does not constrain the magnitude of the top-class softmax probability. A concrete counterexample is a 1D scalar classifier f with logits $(g(\mathbf{x}), 0)$ where $g(0.5) = 10$ and $g(0.6) = 0.01$; the prediction is class 0 on the entire interval $[0.4, 0.6]$ (so f is locally robust at $(0.5, 0)$ at radius 0.1), but the softmax collapses from $(0.99, 0.01)$ at $\mathbf{x} = 0.5$ to $(0.50, 0.50)$ at $\mathbf{x}' = 0.6$. With NLL on a softmax clipped to $[10^{-2}, 1-10^{-2}]$, the source-point bound gives $\ell(f(0.5), 0) \approx 0.01$ and the right-hand side of the naive bound is ≈ 0.01, but the OOD-side loss is $-\log 0.50 \approx 0.69$, a violation by a factor of ≈ 70. The same construction yields a factor-$\approx 2{,}500$ violation for multiclass Brier. The naive source-point assumption therefore controls the loss *at the source* only — not at the shifted point that enters $L^{\ell}_{\mathcal{S}_Q}$.

Repair via Lipschitz augmentation. We restate the theorem with three explicit assumptions: a source-point floor (still measurable from ID data), Lipschitz continuity of the loss in the prediction, and Lipschitz continuity of the network in the input. Together these honestly price in the cost of moving from $\mathbf{x}$ to $\mathbf{x}'$.

Assumption 3 (Confidence floor on certified source points)**.** *There exists $\varepsilon^{\mathrm{ID}}_{\mathrm{floor}} \geq 0$ such that for every $(\mathbf{x}, y)$ at which f_θ is locally robust at radius δ, $\ell(f_\theta(\mathbf{x}), y) \leq \varepsilon^{\mathrm{ID}}_{\mathrm{floor}}$.*

Assumption 4 (Lipschitz loss)**.** *There exists $L_\ell \geq 0$ such that $|\ell(\mathbf{u}, y) - \ell(\mathbf{v}, y)| \leq L_\ell \|\mathbf{u} - \mathbf{v}\|$ for all $\mathbf{u}, \mathbf{v} \in \mathbb{R}^K$ and $y \in \mathcal{Y}$, in some fixed norm on $\mathbb{R}^K$.*

Assumption 5 (Lipschitz network)**.** *There exists $L_f \geq 0$ such that $\|f_\theta(\mathbf{x}) - f_\theta(\mathbf{x}')\| \leq L_f \|\mathbf{x} - \mathbf{x}'\|_\infty$ for all $\mathbf{x}, \mathbf{x}' \in [0, 1]^d$.*

For losses with $\ell(\mathbf{u}, y) = 0$ when $\arg\max_j u_j = y$ (the prediction-induced family covered by Theorem 2), Assumption 3 is satisfied with $\varepsilon^{\mathrm{ID}}_{\mathrm{floor}} = 0$ and the Lipschitz slack $L_\ell L_f \delta$ contracts to 0 as well (since the loss is locally constant on the shift-induced perturbation), recovering Theorem 2 as a special case. For NLL with a softmax clipped to $[\varepsilon_{\mathrm{soft}}, 1 - \varepsilon_{\mathrm{soft}}]$ on a K-class task, $\varepsilon^{\mathrm{ID}}_{\mathrm{floor}} \leq -\log(1 - \varepsilon_{\mathrm{soft}})$ and $L_\ell \leq \varepsilon^{-1}_{\mathrm{soft}}$; for Brier, $\varepsilon^{\mathrm{ID}}_{\mathrm{floor}} \leq K\varepsilon^2_{\mathrm{soft}}$ and $L_\ell \leq 2\sqrt{K}$. The network constant L_f is the global Lipschitz constant of the logits, bounded by the product of layerwise spectral norms (Tsuzuku et al., 2018).

Theorem 3 (CRA-bounded loss shift, Lipschitz-augmented confidence-floor form)**.** *Under Assumptions 1–2 and 3–5, for any bounded loss ℓ with $\ell \le M$ and any $\delta \ge 0$,*

$$L^{\ell}_{\mathcal{S}_Q}(f_\theta) \;\le\; (\varepsilon^{\mathrm{ID}}_{\mathrm{floor}} + L_\ell L_f \delta) \cdot \mathrm{CRA}(f_\theta, \mathcal{S}_P, \delta) \;+\; M \cdot \big[(1 - \mathrm{CRA}(f_\theta, \mathcal{S}_P, \delta)) + \hat{\eta}(\delta)\big]. \quad (16)$$

Proof. Same per-index dichotomy as Theorem 2, but with the GOOD-side bound updated to use Lipschitz continuity. Fix $\delta \ge 0$. On the GOOD event $E_i := \{R_i(\delta)\} \cap \{\|\mathbf{x}'_i - \mathbf{x}_i\|_\infty \le \delta\}$,

$$\begin{aligned} \ell(f_\theta(\mathbf{x}'_i), y_i) &\le \ell(f_\theta(\mathbf{x}_i), y_i) + |\ell(f_\theta(\mathbf{x}'_i), y_i) - \ell(f_\theta(\mathbf{x}_i), y_i)| \\ &\le \varepsilon^{\mathrm{ID}}_{\mathrm{floor}} + L_\ell \, \|f_\theta(\mathbf{x}'_i) - f_\theta(\mathbf{x}_i)\| \\ &\le \varepsilon^{\mathrm{ID}}_{\mathrm{floor}} + L_\ell L_f \delta, \end{aligned}$$

using the triangle inequality, then Assumption 3 on the first summand and Assumption 4 + Assumption 5 on the second. On E_i^c, $\ell(f_\theta(\mathbf{x}'_i), y_i) \le M$. Summing over i as in Theorem 2 and dividing by N gives (16). The Lean theorem `empirical_loss_shift_bound_floor` takes the shifted-point bound (the resulting inequality $\ell(f_\theta(\mathbf{x}'_i), y_i) \le \varepsilon^{\mathrm{ID}}_{\mathrm{floor}} + L_\ell L_f \delta$) as a direct hypothesis; the Lipschitz derivation supplies it. □

ECE. ECE is outside the scope of Theorem 3: it is a binned statistic of the joint sample, not a pointwise loss $\ell : \mathbb{R}^K \times \mathcal{Y} \to [0, M]$, so Assumption 3 does not type-check for it. We report ECE empirically in Section 5.3 and observe a Spearman correlation $\rho \le -0.84$ with ID CRA, which is consistent with the Lipschitz mechanism — high CRA forces sharper softmax distributions on certified points, and the binwise calibration error inherits this structure — but is not a corollary of the bound.

Empirical instantiation. The bound is tight when both $\varepsilon^{\mathrm{ID}}_{\mathrm{floor}}$ and $L_\ell L_f \delta$ are small. We discuss each ingredient separately because, for the architectures we verify, the four constants enter at very different scales.

The source-point floor $\varepsilon^{\mathrm{ID}}_{\mathrm{floor}}$ is observable from the trained model's softmax output on certified ID points. The mean MSP on the full ID test set, reported in Table 17, ranges from 0.876 (TRADES-0.01 on GTSRB) to 0.959 (Plain on GTSRB) for the matched-clean-accuracy families and 0.907–0.934 for MNIST-FC. The certified-conditional MSP is in general higher than the test-set-wide mean (certified points are by definition far from the decision boundary), but our released artifacts do not separately report it; we therefore read $\varepsilon^{\mathrm{ID}}_{\mathrm{floor}}$ as a typical-case quantity. The *uniform* bound that Assumption 3 requires is determined by the worst-confidence certified point, which, for any clipped-softmax implementation with floor $\varepsilon_{\mathrm{soft}}$, is at most $-\log(1 - \varepsilon_{\mathrm{soft}})$ for NLL and $K\,\varepsilon^2_{\mathrm{soft}}$ for Brier.

Softmax clipping. The training pipeline (`torch.nn.functional.cross_entropy` on raw logits) does not itself clip the softmax; clipping is a property of the *evaluator* that computes the loss to which Theorem 3 is applied. Any post-hoc clipping with $\varepsilon_{\mathrm{soft}} > 0$ makes the loss bounded (M finite) and gives a well-defined $\varepsilon^{\mathrm{ID}}_{\mathrm{floor}}$ uniform bound. Without clipping, NLL is unbounded on data where the softmax saturates against the true class, so the theorem applies only to the clipped variant (which is the form universally used for numerical reporting).

The Lipschitz product $L_\ell L_f \delta$ is the load-bearing constant. For NLL viewed as a function of the logits in ℓ_∞ norm, L_ℓ is bounded by 1 ($\partial \mathrm{NLL}/\partial u_j = \mathrm{softmax}(u)_j - \mathbf{1}[j = y]$, which has ℓ_∞-norm ≤ 1); for Brier in ℓ_2, $L_\ell \le 2\sqrt{K}$. The network constant L_f depends strongly on depth: a global upper bound by the product of layerwise spectral norms grows exponentially in the depth of unconstrained networks (Tsuzuku et al., 2018), which can make global bounds loose. We do not estimate L_f empirically. Moreover, the measured natural-shift defect is one at $\delta \in \{1, 2, 3, 4\}/255$ (Appendix D.5), so the loss-shift slack is already at least M there. We therefore do not claim an informative numerical bound for either the shallow baseline or the deeper variants on these axes.

ECE remains outside the framework (it is a binned statistic of the joint sample, not a pointwise loss), and its empirical correlation is a corroboration of the same confidence-stability mechanism rather than a corollary of the bound.

A.8 When the Bound Is Informative

Rank-preservation bound. Theorem 1 guarantees a positive OOD gap when the positive ID gap $\Delta_{\mathcal{S}_P}(\varepsilon)$ exceeds the budget $\alpha_A^P(\varepsilon,\delta) + \alpha_B^P(\varepsilon,\delta) + 2\hat{\eta}(\delta)$. For a fixed coupling, increasing δ decreases the defect but increases both modulus terms; a useful choice minimizes this budget over the admissible displacement radii. A small defect alone is insufficient if the modulus terms consume the ID gap.

On every natural axis measured in Appendix D.5, the defect is one through $16/255$, including all certified radii. The budget is therefore at least two, larger than any possible positive CRA gap. It gives no rank-preservation guarantee on these axes; the agreement and reversals in Section 5.1 are empirical observations. The controlled shifts in Section 5.2 have explicit pairings and known defects; the budget decides the ordering in 9 pair-cells, and all of those orderings hold.

Loss-shift bound. Theorem 2 has an analogous trade-off: increasing δ decreases both ID CRA at radius δ and the defect. Its slack is $M[(1-\mathrm{CRA}(f_\theta,\mathcal{S}_P,\delta)) + \hat{\eta}(\delta)]$. For a nonnegative loss bounded by M, the resulting upper bound improves on the trivial bound only if the ID loss plus this slack is below M. In particular, $\hat{\eta}(\delta) = 1$ makes the slack at least M. The measured natural-shift defects therefore prevent a nontrivial loss guarantee at the evaluated radii, including $\delta \in \{1,2,3,4\}/255$. Corollary 3 also needs its full gap-and-slack condition to hold; nonzero ID CRA alone does not establish an OOD-loss ordering. The correlations in Section 5.3 are measured associations.

Confidence-floor extension. Theorem 3 adds two knobs to the rank-preservation δ-trade-off: an ID-side floor $\varepsilon_{\mathrm{floor}}^{\mathrm{ID}}$ and a Lipschitz-augmented shift cost $L_\ell L_f \delta$. The bound is informative when (i) $\varepsilon_{\mathrm{floor}}^{\mathrm{ID}}$ is small (i.e., the model is confident-correct on certified source points), (ii) the loss-shift slack $M \cdot [(1-\mathrm{CRA}) + \hat{\eta}]$ is small, and (iii) the Lipschitz product $L_\ell L_f \delta$ is small at the chosen δ. In our experiments, $\varepsilon_{\mathrm{floor}}^{\mathrm{ID}}$ is dominated by the MSP at certified points, which adversarial training systematically pushes toward 1 (Appendix E.2). The Lipschitz-loss constants are loss-specific and known: $L_\ell \le \varepsilon_{\mathrm{soft}}^{-1}$ for clipped-softmax NLL, $L_\ell \le 2\sqrt{K}$ for Brier. The Lipschitz-network constant L_f must also be bounded to quantify the extension. We do not estimate it empirically. Moreover, the measured natural-shift defect of one already makes the loss-shift slack at least M at the evaluated radii, so we do not claim a nontrivial numerical confidence-floor bound for these axes.

A.9 Relation to Generalization Bounds

The CRA-bounded loss-shift result (Theorem 2) connects to the broader generalization-under-shift literature. Domain-adaptation bounds based on $\mathcal{H}\Delta\mathcal{H}$-divergence (Ben-David et al., 2010) and discrepancy (Mansour et al., 2009) produce additive shift terms structurally similar to ours, but use a hypothesis-class divergence rather than a per-input certificate; PAC-Bayes margin bounds (Neyshabur et al., 2018) replace the certified radius with a learnt-distribution proxy, and Wasserstein-DRO frameworks (Sinha et al., 2018; Kumar et al., 2023) bound the worst-case loss over a distribution ball. Our bound trades the breadth of those frameworks for a directly observable per-input quantity, the certified radius, and is therefore deterministic rather than concentration-based or worst-case.

B Verifier-Reported vs. True Certified Radius

For a sound verifier, the reported lower bound $\hat{\rho}_{f_\theta}$ on the certified radius satisfies $\hat{\rho}_{f_\theta}(\mathbf{x},y) \le \rho_{f_\theta}(\mathbf{x},y)$ for every input $(\mathbf{x},y)$. Whenever the verifier returns `safe` on $\mathcal{B}_\infty(\mathbf{x},\varepsilon)$, the local-robustness condition of Definition 1 holds at radius ε, so $\rho_{f_\theta}(\mathbf{x},y) \ge \varepsilon$. Conversely, an `unsafe` or `unknown` verdict does not imply $\rho_{f_\theta}(\mathbf{x},y) < \varepsilon$; it only means the verifier could not produce a proof at that radius. Throughout the main paper we report CRA computed from `safe` verdicts, which is therefore a sound (and in general non-tight) lower bound on the population-level certified accuracy defined directly in terms of ρ_{f_θ}. All comparisons across models use the same verifier (α,β-CROWN) and the same property set per distribution, so the gap $\rho_{f_\theta} - \hat{\rho}_{f_\theta}$ is held as constant as possible across conditions.

C Experimental Details

C.1 Architecture Specifications

This section gives the full layer-by-layer specification of every model used in this paper, together with parameter counts. All models accept raw $[0, 1]$ pixel values without input normalization, matching the convention expected by α, β-CROWN at verification time. Each model is implemented as a `torch.nn.Sequential` module so that the verifier can ingest it directly via ONNX export with no architectural surgery.

C.1.1 MNIST-FC Family

All MNIST-FC models take a $1\times28\times28$ input, flatten it to $\mathbb{R}^{784}$, pass through a stack of `Linear`→`ReLU` blocks, and emit a 10-class logit vector. The first model (`mnist_fc`) is the main-experiment baseline used in Section 5.1; the remaining four are the architecture variants pooled in Appendix E.1 (all but 3×512).

Table 4: MNIST-FC architectures. "Hidden dims" lists the widths of the hidden layers in order; each is followed by ReLU, except the final classifier. Input is $1\times28\times28$ flattened to 784; output is 10 logits.

Function	Role	Hidden dims	# Hidden layers	Params
`mnist_fc`	Main baseline (Sec. 5.1)	$10-10$	2	8,070
`mnist_fc_2_200`	Shallow ablation	200	1	159,010
`mnist_6_100`	Narrow-Deep ablation	$100^{\times5}$	5	119,910
`mnist_9_100`	Deeper-Narrow ablation	$100^{\times8}$	8	150,210
`mnist_fc_3_512`	Wide-Shallow ablation	$512^{\times3}$	3	932,362

Layer-by-layer specification (`mnist_fc`).

```
Sequential(
  (0):  Flatten()                                    # 1x28x28 -> 784
  (1):  Linear(in=784, out=10, bias=True)
  (2):  ReLU()
  (3):  Linear(in=10,  out=10, bias=True)
  (4):  ReLU()
  (5):  Linear(in=10,  out=10, bias=True)            # 10-class logits
)
```

The remaining four MNIST-FC variants follow the same template: `Flatten` → (N_h hidden `Linear`→`ReLU` blocks) → `Linear(·, 10)`, with hidden dims as listed in Table 4. None of the MNIST-FC variants use batch normalization, dropout, or any non-linearity other than ReLU.

C.1.2 GTSRB-CNN Family

The GTSRB-CNN models take a $3\times32\times32$ image and produce 43 logits (one per class). The convolutional stack downsamples $32\times32 \to 16\times16 \to 8\times8$ via two stride-2 convolutions, followed by a 512-d fully connected hidden layer and a 43-way classifier. Section 5.1 uses `gtsrb_cnn`.

Table 5: GTSRB-CNN architecture. Input is $3\times32\times32$; output is 43 logits.

Function	Role	Conv widths	Params
`gtsrb_cnn`	Baseline (no normalization, Sec. 5.1)	32, 32, 64, 64	2,221,131

Layer-by-layer specification (`gtsrb_cnn`).

```
Sequential(
  (0):  Conv2d(3,  32, k=3, stride=1, padding=1)     # 32x32x32
  (1):  ReLU()
```

```
  (2):  Conv2d(32, 32, k=4, stride=2, padding=1)      # 16x16x32
  (3):  ReLU()
  (4):  Conv2d(32, 64, k=3, stride=1, padding=1)      # 16x16x64
  (5):  ReLU()
  (6):  Conv2d(64, 64, k=4, stride=2, padding=1)      #  8x8x64
  (7):  ReLU()
  (8):  Flatten()                                     # 4096
  (9):  Linear(in=4096, out=512, bias=True)
  (10): ReLU()
  (11): Linear(in=512,  out=43, bias=True)            # 43-class logits
)
```

C.1.3 CIFAR-CNN FAMILY

The CIFAR-CNN models take a $3\times32\times32$ image and produce 10 logits. The network is the `cifar_conv_small` architecture distributed with α, β-CROWN: two convolutions with 4×4 kernels and stride 2 (16 and 32 channels, no padding), each followed by ReLU, then a fully connected layer of width 100 with ReLU and a linear output layer.

C.2 MODEL TRAINING CONFIGURATIONS

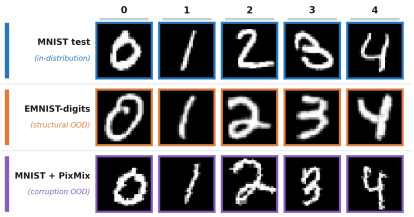
Table 6 summarizes the per-model training procedure, training ε (where applicable), and clean test accuracy for all three families.

Table 6: Model training configurations and in-distribution clean accuracy. All models within each architecture are matched on clean accuracy. The GTSRB plain model uses a minimal TRADES configuration; undefended GTSRB training yields near-zero CRA.

Architecture	Model	Training	$\varepsilon_{\text{train}}$	Clean Acc.
MNIST-FC	Plain	CE	—	93.2%
	Adv-0.01	PGD-AT	0.01	93.2%
	Adv-0.02	PGD-AT	0.02	92.6%
	Adv-0.03	PGD-AT	0.03	92.6%
GTSRB-CNN	Plain	TRADES (β=6.0)	0.01	89.3%
	TRADES-0.01	TRADES (β=6.0)	0.01	89.0%
	TRADES-0.02	TRADES (β=3.0)	0.02	87.3%†
	TRADES-0.03	TRADES (β=3.0)	0.03	88.1%
CIFAR-CNN	Plain	CE	—	58.4%
	Adv-0.5	PGD-AT	0.5/255	58.6%
	Adv-1	PGD-AT	1/255	58.8%
	Adv-1.5	PGD-AT	1.5/255	58.7%
	Adv-2	PGD-AT	2/255	58.1%
	Adv-2.5	PGD-AT	2.5/255	58.3%
	Adv-3	PGD-AT	3/255	58.1%

†TRADES-0.02 achieves slightly lower clean accuracy despite achieving the highest in-distribution CRA; this trade-off is expected at larger $\varepsilon_{\text{train}}$. CIFAR-CNN models are trained from scratch with Adam (learning rate 10^{-3}, batch 128); PGD-AT uses 7 steps of size $2/255$; every model is stopped at the first epoch whose clean test accuracy reaches 58%, which the weakly regularized variants reach early (Adv-0.5 at epoch 9).

Compute resources. All experiments were carried out on a single workstation with one NVIDIA RTX A6000 GPU (48 GB) and a 16-core AMD CPU. Training each MNIST-FC or GTSRB-CNN model takes well under 30 minutes; the full 17-model architecture ablation completes in a few GPU-hours total. Verification with α, β-CROWN, Marabou, and NNV uses a 60-second per-instance timeout (Appendix F.1), so the per-distribution wall time is bounded by $N_{\text{properties}} \cdot 60\,\text{s}$ per model, but most properties complete in well under a second. Total verification compute for all (model, distribution, ε) cells reported in this paper is on the order of a few hundred CPU-hours.



(a) MNIST-FC. Columns are digit classes 0–4. Rows: **MNIST test**, **EMNIST-digits**, **MNIST+PixMix**.



(b) GTSRB-CNN. Columns are sign categories (name at top). Rows: **GTSRB test**, **BTSD**, **GTSRB+PixMix**.

Figure 6: Representative images from the three evaluation distributions for each task. Rows correspond to in-distribution, structural OOD, and corruption OOD sources (colored bar and label at left). EMNIST images are orientation-corrected to match MNIST's canonical layout; BTSD images have higher contrast and simpler backgrounds than GTSRB.

C.3 Probabilistic OOD Metric Definitions

For a model producing softmax probabilities $p_{i,c}$ on test sample i for class $c \in \{1, \ldots, C\}$ with true label y_i, predicted class $\hat{y}_i = \arg\max_c p_{i,c}$, and N samples, the six metrics used in Section 5.3 and Appendix E.2 are defined as follows.

- **Macro-F1** averages per-class F_1 scores uniformly: $\text{macro-F}_1 = \frac{1}{C}\sum_c F_{1,c}$. Treats every class equally regardless of frequency; sensitive to class-imbalanced errors that accuracy ignores.
- **Expected Calibration Error (ECE)** with $B = 15$ equal-width bins on top-label confidence: $\text{ECE} = \sum_b \frac{|S_b|}{N} \left|\text{acc}(S_b) - \text{conf}(S_b)\right|$, where S_b is the set of samples whose top-class probability falls in bin b. Measures the gap between the model's confidence and its accuracy.
- **Brier score (multiclass)**: $\text{Brier} = \frac{1}{N}\sum_i \sum_c \left(p_{i,c} - \mathbf{1}[y_i = c]\right)^2$. A proper scoring rule that penalizes the full softmax distribution against the one-hot ground truth; decomposes into calibration plus refinement.
- **Negative log-likelihood (NLL)**: $\text{NLL} = -\frac{1}{N}\sum_i \log p_{i,y_i}$. Heavily penalizes confidently-wrong predictions; standard in calibration literature and minimized in expectation by the true conditional distribution.
- **Mean maximum-softmax probability (mean-MSP)**: $\overline{\text{MSP}} = \frac{1}{N}\sum_i \max_c p_{i,c}$. A behavioral probe rather than a quality metric; the diagnostic comparison is $\overline{\text{MSP}}(\text{ID}) > \overline{\text{MSP}}(\text{OOD})$, which a well-calibrated model should satisfy.
- **AUROC vs. ID** using $\max_c p_{i,c}$ as the in-distribution score and labeling ID samples as positives, OOD as negatives. Pure rank-based separation, invariant to score scaling.

D More Detailed Analysis

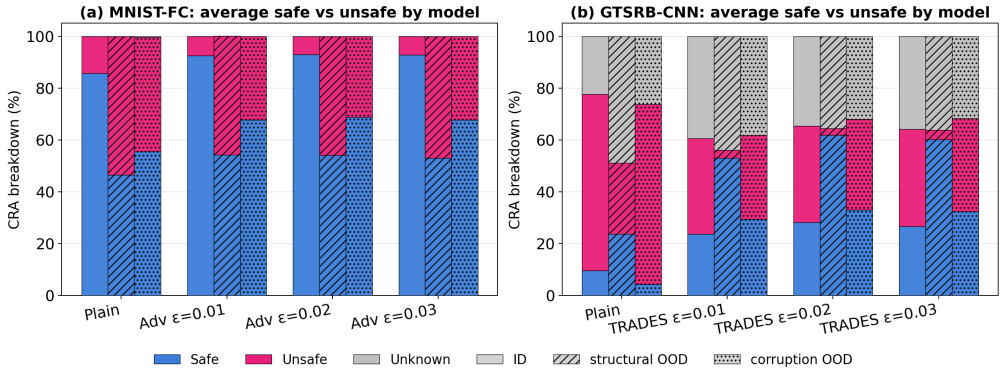
This appendix gives the per-distribution, per-pair and per-family detail behind Experiments 1 and 2 (Sections 5.1 and 5.2).

D.1 Per-Distribution Breakdown

This section expands the condensed per-distribution discussion in Section 5.1 with the granular by-suite breakdowns that were elided from the main body for space. We retain the same model sets and verification configurations as Tables 19 and 16.

D.1.1 MNIST-FC Per-Distribution Breakdown

In-distribution CRA. All three adversarially trained models substantially outperform the plain baseline (85.75% average CRA vs. 92.50–93.00%). Adv-0.02 achieves the highest average CRA at



(a) Verifier outcomes per pair (model, distribution): Safe, Unsafe, and Unknown; bars sum to 100%, blue labels report Avg_CRA.

93.00%, marginally ahead of Adv-0.03 (92.75%) and Adv-0.01 (92.50%). At the largest evaluation epsilon ($\varepsilon = 4/255$), Adv-0.02 and Adv-0.03 both achieve 92.0%, while Adv-0.01 reaches 90.0%.

Structural OOD (EMNIST). CRA drops substantially for all models when evaluated on EMNIST-digits: the plain model falls from 85.75% to 46.50% (a 39.25 percentage point drop), and the best adversarially trained model (Adv-0.01) falls from 92.50% to 54.25% (38.25 pp). Despite these large absolute drops, the rank ordering is fully preserved: Plain $<$ Adv-0.01 $\approx$ Adv-0.02 $\approx$ Adv-0.03.

Corruption OOD (PixMix-MNIST). PixMix yields smaller absolute drops than EMNIST for all models. The plain model achieves 55.50% on PixMix vs. 46.50% on EMNIST; Adv-0.02 achieves 68.75% on PixMix vs. 54.00% on EMNIST. This suggests that PixMix-corrupted MNIST images, while visually degraded, produce inputs whose ℓ_∞ neighborhoods are structurally closer to the MNIST training distribution than EMNIST images are. Rank ordering is again perfectly preserved.

Plain-vs-adversarial separation. The gap between plain and adversarially trained models is *larger* on PixMix OOD than in-distribution, suggesting corruption OOD better reveals the underlying robustness structure than clean test accuracy would indicate.

D.1.2 GTSRB-CNN Per-Distribution Breakdown

In-distribution CRA. The plain GTSRB model achieves only 9.53% average CRA and 0% at $\varepsilon = 4/255$, demonstrating that the undefended network has nearly no certified robustness at these perturbation levels despite its 89.3% clean accuracy. TRADES training dramatically improves CRA: TRADES-0.02 achieves 28.14% average CRA and 4.19% at $\varepsilon = 4/255$. Among the TRADES models, the ranking is TRADES-0.01 $<$ TRADES-0.03 $<$ TRADES-0.02, consistent with the idea that larger training epsilons yield higher CRA up to a point, with TRADES-0.02 providing the best trade-off.

Structural OOD (BTSD). The most striking result is that *all* models achieve higher CRA on BTSD than on the in-distribution GTSRB test set: the plain model improves from 9.53% to 23.73%, and TRADES-0.02 improves from 28.14% to 61.86%. This appears to contradict the expected direction of OOD degradation. Crucially, however, the rank ordering is still fully preserved: Plain $<$ TRADES-0.01 $<$ TRADES-0.03 $<$ TRADES-0.02. We attribute this anomaly to the structural simplicity of BTSD verification queries (higher contrast, simpler backgrounds, fewer adjacent classes per scene) compared to in-distribution GTSRB images.

Corruption OOD (PixMix-GTSRB). For GTSRB+PixMix, CRA drops below in-distribution levels for the plain model (4.30% vs. 9.53%), but rises above in-distribution for all TRADES models (e.g., 33.02% vs. 28.14% for TRADES-0.02). This is consistent with the larger separation that adversarial training carves out around training-distribution inputs. Rank ordering is preserved across all models.

CRA at $\varepsilon = 4/255$. At the largest evaluation epsilon, the plain model achieves 0% CRA on all three distributions: formal verification at this perturbation level is entirely infeasible for the undefended model. Adversarially trained models maintain non-trivial CRA (2.79–4.19% in-distribution; 6.05–13.56% on BTSD), demonstrating that adversarial training genuinely enables formal verification at larger radii.

Figures 8 and 9 show CRA against ε for every distribution condition.

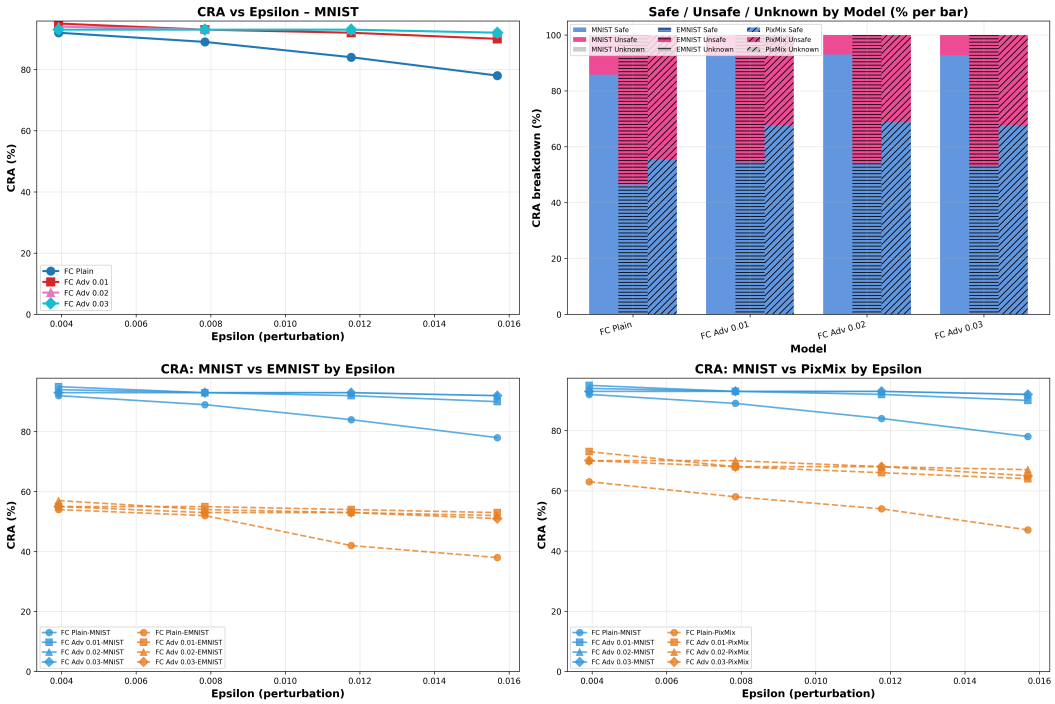


Figure 8: CRA vs. ε for MNIST-FC models across distribution conditions. Top-left: in-distribution (MNIST). Bottom-left: structural OOD (EMNIST). Bottom-right: corruption OOD (PixMix).

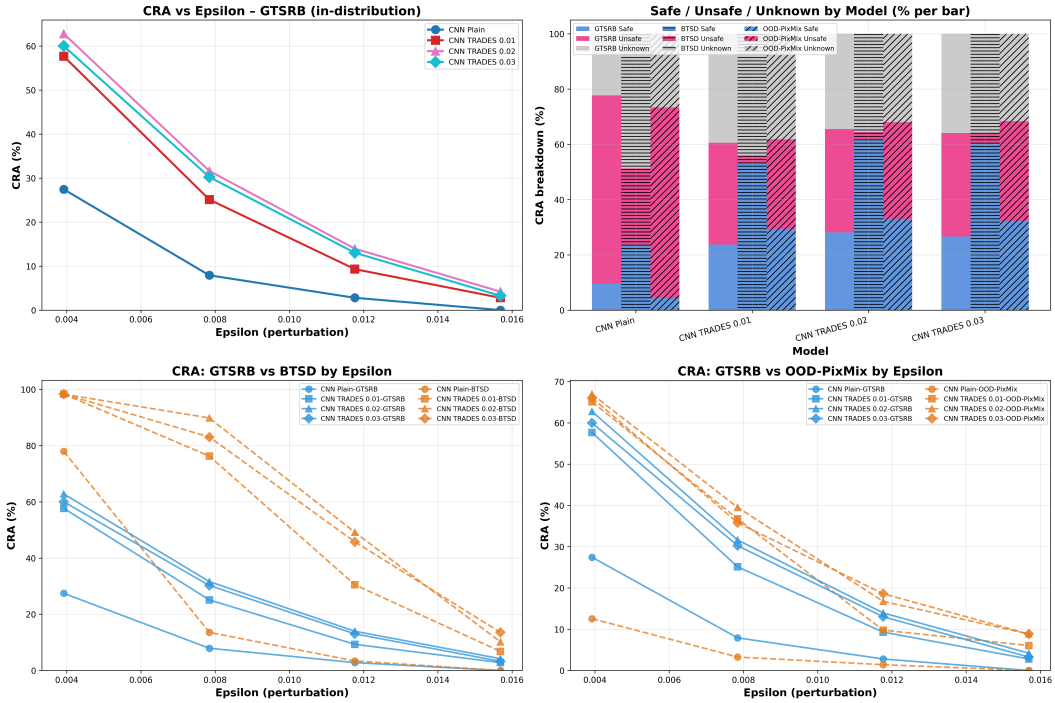


Figure 9: CRA vs. ε for GTSRB-CNN models across distribution conditions. Rows show in-distribution (GTSRB), structural OOD (BTSD), and corruption OOD (PixMix).

D.2 Pairwise Rank Preservation

Section 5.1 compares models pairwise on the property sets for which per-instance verdicts were recorded: for MNIST-FC at $\varepsilon = 2/255$, 100 properties on each of MNIST, EMNIST and PixMix; for GTSRB-CNN at $\varepsilon = 2/255$, 150 on GTSRB, 59 on BTSD and 150 on PixMix; for the seven-model CIFAR family and the TRADES CIFAR family at $\varepsilon \in \{1, 2\}/255$, 100 on CIFAR-10 and 100 (92 for TRADES) on CIFAR-10-C. For each pair the ID leader f is the model with the higher ID CRA; pairs with exactly equal ID CRA (8 of the $70 + 8$) have no order to preserve and are excluded. The ID gap is tested with an exact McNemar test on the properties where the two verdicts differ; the slack is the leader's shift drop minus the other model's, so that $\Delta_{\mathcal{S}_Q} = \Delta_{\mathcal{S}_P} - s$ holds by definition. Every pair, with its gaps, p-values and slack, is in `pairwise_gaps.csv`, generated by `make_iclr_tables.py`; Figure 2 plots them and Table 7 lists the 6 inversions. None involves a significant ID gap, and in none is the OOD gap significant. The two largest slacks belong to leaders whose CRA falls fastest with radius: at $1/255$ Plain's ID CRA drops from 87 to 53 between $1/255$ and $2/255$ against 86 to 74 for Adv-2.5, and under corruption Plain loses 11 points while Adv-2.5 loses none.

Table 7: All rank inversions on the per-instance property sets. $\Delta_{\mathcal{S}_P}$ and $\Delta_{\mathcal{S}_Q}$ in CRA points, with exact McNemar p-values on the shared properties; slack is the ID leader's shift drop minus the other model's.

Family	OOD set	ε	ID leader f	g	$\Delta_{\mathcal{S}_P}$	p_{ID}	$\Delta_{\mathcal{S}_Q}$	p_{OOD}	slack
CIFAR-CNN	CIFAR-10-C	$1/255$	Plain	Adv-0.5	+1	1.00	-2	0.85	+3
CIFAR-CNN	CIFAR-10-C	$1/255$	Plain	Adv-2.5	+1	1.00	-10	0.09	+11
CIFAR-CNN	CIFAR-10-C	$1/255$	Adv-2	Adv-1	+1	1.00	-1	1.00	+2
CIFAR-CNN	CIFAR-10-C	$2/255$	Adv-1.5	Adv-3	+3	0.58	-7	0.14	+10
CIFAR-CNN (TRADES)	CIFAR-10-C	$1/255$	TRADES-0.01/β3	TRADES-0.01/β1	+1	1.00	-3	0.58	+4
CIFAR-CNN (TRADES)	CIFAR-10-C	$2/255$	TRADES-2/255/β6	TRADES-0.01/β3	+5	0.42	-4	0.57	+9

D.3 CIFAR-10: PGD-AT Families and a TRADES Family

Table 8: CRA (%) for the seven-model CIFAR-CNN family at $\varepsilon \in \{1, 2\}/255$ on CIFAR-10 and on the CIFAR-10-C corruption set (N=100 consensus-correct properties per set; α, β-CROWN). **Bold** marks the best model per column. Per-cell standard error is $\leq$ 5pp.

Model	Clean acc.	CIFAR-10 (in-dist.)		CIFAR-10-C (corrupt. OOD)	
		$1/255$	$2/255$	$1/255$	$2/255$
Plain	58.4	87	53	76	42
Adv-0.5	58.6	86	70	78	55
Adv-1	58.8	94	73	91	68
Adv-1.5	58.7	**96**	**89**	93	80
Adv-2	58.1	95	80	90	77
Adv-2.5	58.3	86	74	86	76
Adv-3	58.1	**96**	86	**94**	**87**

The four-model subfamily Plain, Adv-1, Adv-2, Adv-3 mirrors the MNIST-FC and GTSRB-CNN families (one plain model and three adversarially trained ones at increasing budgets). Within it, CRA is monotone in $\varepsilon_{\text{train}}$ on both sets (Table 9), and the ID and CIFAR-10-C orderings agree with Spearman $\rho = 1.00$ ($[0.80, 1.00]$) at $\varepsilon = 2/255$ and $\rho = 0.80$ ($[0.20, 1.00]$) at $\varepsilon = 1/255$, where Adv-1 and Adv-2 are within one point of each other.

Table 9: CRA (%) for the four-model CIFAR-CNN subfamily (N=100 per set).

Model	Clean acc.	CIFAR-10 (in-dist.)		CIFAR-10-C (corrupt. OOD)	
		$1/255$	$2/255$	$1/255$	$2/255$
Plain	58.4	87	53	76	42
Adv-1	58.8	94	73	91	68
Adv-2	58.1	95	80	90	77
Adv-3	58.1	**96**	**86**	**94**	**87**

A TRADES family. The GTSRB-CNN family is trained with TRADES; to check that the CIFAR result does not depend on the objective, we train a second CIFAR family with the same TRADES code (10-step perturbation, CE plus β times the KL term). The GTSRB budgets themselves ($\varepsilon \in \{0.01, 0.02, 0.03\}$ with $\beta \in \{6, 3, 3\}$) reduce clean accuracy on this network to 46.7%, 39.8% and 38.9%, too low to match; the configurations that converge close together are $\varepsilon = 2/255$ with $\beta = 6$ (52.0%), and $\varepsilon = 0.01$ with $\beta = 3$ (51.9%) and $\beta = 1$ (53.8%), with a Plain model stopped at 53.5%, a 1.9-point spread, matched by configuration as on GTSRB. Properties are generated as for the other families (100 consensus-correct ID points; 92 on CIFAR-10-C, where this family agrees correctly on fewer images). Table 10 gives the result. Plain is lowest on every set and significantly so under paired tests at $\varepsilon = 2/255$ in all six comparisons (McNemar $p \leq 0.001$; gaps of 21–32 points) and in four of six at $1/255$; no two robust models are distinguishable on any set ($p \geq 0.14$). Rank agreement between ID and CIFAR-10-C is Spearman $\rho = 0.80$ ($[0.20, 1.00]$) at $\varepsilon = 2/255$ and 0.80 ($[0.11, 1.00]$) at $1/255$, with the only order changes inside the tied robust triple, and the model with the highest ID CRA is the OOD-best model at both radii (zero regret, Table 18). One detail is worth recording: on this network the smallest KL weight, $\beta = 1$, gives the highest CRA at $\varepsilon = 2/255$.

Table 10: CRA (%) for the four-model TRADES CIFAR-CNN family (N=100 ID and N=92 CIFAR-10-C properties).

Model	Clean acc.	CIFAR-10 (in-dist.)		CIFAR-10-C (corrupt. OOD)	
		$^1/_{255}$	$^2/_{255}$	$^1/_{255}$	$^2/_{255}$
Plain	53.5	84	51	80	49
TRADES-0.01/β3	51.9	94	77	90	74
TRADES-2/255/β6	52.0	**95**	82	**95**	70
TRADES-0.01/β1	53.8	93	**83**	93	**79**

D.4 Controlled Shift and Partial-Defect Sweep

The natural shift axes all have $\hat{\eta}(\delta) = 1$ at every certified radius (Appendix D.5), so the budget condition (10) cannot be met on them. To test the bounds where they apply, we construct shifts with a known defect. **δ-shift.** Starting from 100 MNIST test points on which all four MNIST-FC models are correct, each point is moved by a random sign pattern of magnitude exactly δ in every coordinate, clipped to $[0, 1]^d$, for $\delta \in \{1, 2, 4, 8, 16\}/255$; the identity coupling then has $\hat{\eta}(\delta) = 0$, and Lemma 2 predicts $\mathrm{CRA}(f, \mathcal{S}_Q, \varepsilon) \geq \mathrm{CRA}(f, \mathcal{S}_P, \varepsilon + \delta)$. **Mixture.** A fraction $\hat{\eta}$ of the points is instead moved by a larger δ_{large}, so that the defect at δ equals $\hat{\eta}$ by construction and (5) predicts $\mathrm{CRA}(f, \mathcal{S}_P, \varepsilon+\delta) - \hat{\eta}$. The shifted boxes are verified with α, β-CROWN exactly as the in-distribution properties are.

Table 11 reports the δ-shift results at $\varepsilon = 2/255$. Every model is verified on the same 100 properties, so the comparison between models is paired: the last three columns give the per-property difference between Plain and each robust model with a paired bootstrap interval, starred when an exact McNemar test on the discordant properties gives $p < 0.05$. Plain has the lowest CRA at every δ and is significantly lower than each robust model for $\delta \geq 4/255$. The three robust models are not distinguished by the data at any δ; the one nominal change in their order (Adv-0.01 at 97 against Adv-0.03 at 99 for $\delta = 8/255$) has a paired interval of $[-6, +2]$ points and $p = 0.63$. In all 20 cells the verifier decided every instance (Table 12), so the reported values equal the verifier's decision on every point.

Table 11: Controlled δ-shift on MNIST-FC at $\varepsilon = 2/255$ ($n = 100$ shared properties, $\hat{\eta}(\delta) = 0$). Left: CRA per model. Right: paired difference in CRA between Plain and each robust model, with a paired-bootstrap 95% interval; * marks McNemar $p < 0.05$.

δ	Plain	Adv-0.01	Adv-0.02	Adv-0.03	Plain−Adv-0.01	Plain−Adv-0.02	Plain−Adv-0.03
$1/255$	94	99	99	98	$-5\,[-10, -1]$	$-5\,[-10, -1]$	$-4\,[-10, +1]$
$2/255$	94	99	99	98	$-5\,[-10, -1]$	$-5\,[-10, -1]$	$-4\,[-10, +1]$
$4/255$	89	99	99	99	$-10\,[-16, -4]^*$	$-10\,[-16, -4]^*$	$-10\,[-17, -4]^*$
$8/255$	85	97	99	99	$-12\,[-19, -6]^*$	$-14\,[-21, -8]^*$	$-14\,[-22, -7]^*$
$16/255$	69	98	98	97	$-29\,[-38, -21]^*$	$-29\,[-38, -20]^*$	$-28\,[-37, -19]^*$

Table 13 summarizes the full sweep: 320 verified cells over 8 models in two architecture families (the four MNIST-FC models of the main text and a 2×200 family that includes a CROWN-IBP-trained

Table 12: Number of instances (of 100) left undecided by the verifier within the 60-second timeout in each controlled-shift cell at $\varepsilon = 2/255$.

Model	$\delta = 1/255$	$\delta = 2/255$	$\delta = 4/255$	$\delta = 8/255$	$\delta = 16/255$
Plain	0	0	0	0	0
Adv-0.01	0	0	0	0	0
Adv-0.02	0	0	0	0	0
Adv-0.03	0	0	0	0	0

member), $\varepsilon \in \{1, 2, 4\}/255$, both constructions, and $\hat{\eta} \in \{0, 0.25, 0.5, 0.75, 1\}$. The lower bound is never violated. It is tight when the defect is zero, with a median slack of a few points, and loosens as $\hat{\eta}$ grows, because the defect term charges a full point for every moved input whether or not the move crosses the decision boundary. Table 14 shows this for the MNIST-FC models at $\varepsilon = 2/255$.

Table 13: Partial-defect sweep: for each construction, family and radius, the number of verified cells, the number in which measured CRA falls below the predicted lower bound, and the median and maximum slack (measured minus predicted).

Construction	Family	ε	Cells	Bound violations	Median slack (pp)	Max slack (pp)
δ-shift ($\hat{\eta} = 0$)	FC-2x200	$2/255$	20	0	1.0	89
δ-shift ($\hat{\eta} = 0$)	MNIST-FC	$1/255$	20	0	4.0	83
δ-shift ($\hat{\eta} = 0$)	MNIST-FC	$2/255$	20	0	5.5	72
δ-shift ($\hat{\eta} = 0$)	MNIST-FC	$4/255$	20	0	12.0	61
mixture ($\hat{\eta} \in [0, 1]$)	FC-2x200	$2/255$	80	0	5.2	89
mixture ($\hat{\eta} \in [0, 1]$)	MNIST-FC	$1/255$	40	0	8.0	83
mixture ($\hat{\eta} \in [0, 1]$)	MNIST-FC	$2/255$	80	0	14.2	72
mixture ($\hat{\eta} \in [0, 1]$)	MNIST-FC	$4/255$	40	0	16.4	61

Table 14: Measured CRA [predicted lower bound] for the mixture construction on MNIST-FC at $\varepsilon = 2/255$, $\delta = 1/255$, $\delta_{\text{large}} = 8/255$, as the defect $\hat{\eta}$ increases.

Model	$\hat{\eta} = 0$	$\hat{\eta} = 0.25$	$\hat{\eta} = 0.5$	$\hat{\eta} = 0.75$	$\hat{\eta} = 1$
Plain	94 [93]	92 [73]	90 [53]	87 [33]	85 [13]
Adv-0.01	99 [99]	98 [96]	98 [94]	98 [92]	97 [89]
Adv-0.02	99 [99]	99 [98]	99 [96]	99 [94]	99 [93]
Adv-0.03	98 [98]	98 [94]	98 [91]	99 [88]	99 [84]

D.5 Measured Shift Defect on the Natural Axes

Table 15 reports $\hat{\eta}(\delta)$ in ℓ_∞ for the five natural shift axes, from 400 pairs per axis. For the corruption axes the pipeline that produced the PixMix images is replayed from the same source images, so each corrupted image is paired with its own source and the identity coupling is exact; the full transformation is decomposed into its corruption-only and geometric-only components. For the structural axes, and for CIFAR-10-C, whose pooled subset lost its correspondence to the clean images, no per-image pairing exists; each shifted image is paired with its nearest same-class source image in ℓ_∞, and the resulting distance is a lower bound on the transport that any label-preserving coupling must pay for that image. Because the defect is non-increasing in δ, a lower bound on every displacement gives a lower bound on $\hat{\eta}(\delta)$; the value 1.00 is therefore exact.

At every certified radius used in the paper ($\delta \leq 4/255$), and up to $16/255$, every natural axis has $\hat{\eta}(\delta) = 1.00$: no pair is displaced by less than the largest certified radius. Even the corruption-only component of PixMix moves every image by more than $16/255$ in ℓ_∞. The budget condition (10) is therefore not met on any natural axis, and Theorem 1 makes no prediction there; the rank agreement reported in Section 5 on these axes is measured, not guaranteed. The geometric-only component of the GTSRB corruption is the only case with $\hat{\eta} < 1$, and it is not a shift the paper evaluates on its own.

Table 15: Empirical shift defect $\hat{\eta}(\delta)$ in ℓ_∞ on the five natural axes ($n = 400$ pairs per axis). Identity couplings are exact; nearest-neighbour distances lower-bound the transport of any label-preserving coupling on these samples.

Axis	Coupling	Variant	$\hat{\eta}(1/255)$	$\hat{\eta}(4/255)$	$\hat{\eta}(16/255)$
MNIST → PixMix	identity (exact)	full	1.00	1.00	1.00
MNIST → PixMix	identity (exact)	corruption only	1.00	1.00	1.00
MNIST → PixMix	identity (exact)	geometric only	1.00	1.00	1.00
GTSRB → PixMix	identity (exact)	full	1.00	1.00	1.00
GTSRB → PixMix	identity (exact)	corruption only	1.00	1.00	1.00
GTSRB → PixMix	identity (exact)	geometric only	0.73	0.73	0.66
MNIST → EMNIST	NN distance (lower bound)	–	1.00	1.00	1.00
GTSRB → BTSD	NN distance (lower bound)	–	1.00	1.00	1.00
CIFAR-10 → CIFAR-10-C	NN distance (lower bound)	–	1.00	1.00	1.00

Table 16: CRA (%) for GTSRB-CNN models at each $\varepsilon \in \{1, 2, 3, 4\}/255$ and mean (*Avg*), across three test distributions, verified with α, β-CROWN. GTSRB: 215 properties (5 per class); BTSD: 59 properties; PixMix: 215 properties. The standard error on a single per-ε cell follows $\sqrt{p(1-p)/N}$ ($\leq$ 3.4pp on GTSRB/PixMix, $\leq$ 6.5pp on BTSD); claims target rank order rather than absolute magnitudes. **Bold** marks the best model per column. †Plain model has 37–57 unknown/timeout instances per ε on GTSRB.

	GTSRB (in-dist.)					BTSD (struct. OOD)					PixMix (corrupt. OOD)				
Model	$\frac{1}{255}$	$\frac{2}{255}$	$\frac{3}{255}$	$\frac{4}{255}$	Avg	$\frac{1}{255}$	$\frac{2}{255}$	$\frac{3}{255}$	$\frac{4}{255}$	Avg	$\frac{1}{255}$	$\frac{2}{255}$	$\frac{3}{255}$	$\frac{4}{255}$	Avg
Plain†	27.44	7.91	2.79	0.00	9.53	77.97	13.56	3.39	0.00	23.73	12.56	3.26	1.40	0.00	4.30
TRADES-0.01	57.67	25.12	9.30	2.79	23.72	**98.31**	76.27	30.51	6.78	52.97	65.12	36.74	9.77	6.05	29.42
TRADES-0.02	**62.79**	**31.63**	**13.95**	**4.19**	**28.14**	**98.31**	**89.83**	**49.15**	10.17	**61.86**	**66.98**	**39.53**	16.74	**8.84**	**33.02**
TRADES-0.03	60.00	30.23	13.02	3.26	26.63	**98.31**	83.05	45.76	**13.56**	60.17	66.05	35.81	**18.60**	**8.84**	32.33

E Additional Ablation Studies

This appendix collects the cross-architecture ID/OOD CRA correlation, the probabilistic OOD metrics in tabular form, and the model-selection regret analysis behind Experiment 3 (Section 5.3).

E.1 ID CRA Predicts OOD CRA

For the correlation analysis we pool four MNIST-FC architectures (`mnist_fc`, `mnist_fc_2_200`, `mnist_6_100`, `mnist_9_100`) with their four-to-five training-ε settings (plain, Adv-0.01/0.02/0.03, plus a CROWN-IBP-trained variant at $\varepsilon = 0.1$), yielding 17 MNIST classifiers verified on identical 100-property sets per suite.

Findings. ID CRA correlates strongly with OOD CRA on both shifted suites. Across the 17-model summary (one CRA value per model, averaged over ε), Spearman $\rho = +0.80$ on EMNIST and $\rho = +0.87$ on PixMix-MNIST: the rank ordering imposed by ID CRA carries over to the OOD suites. Across all (model, ε) pairs the correlation remains tight in Pearson terms ($r = 0.667$ for EMNIST, $r = 0.827$ for PixMix); the correlation is somewhat tighter for PixMix (corruption shift) than for EMNIST (structural shift), consistent with PixMix being a less severe distributional change. All 17 points lie below $y = x$ — absolute CRA always drops under shift — but the relative ordering is preserved. Adversarially trained models (colored markers) cluster in the upper-right while plain models (gray) concentrate in the lower-left, showing that training variant is the primary driver of CRA level across both distributions. The epsilon trajectories (thin lines) are approximately parallel, indicating that the ID–OOD relationship is stable across perturbation radii. Together with the per-architecture rank-preservation evidence in Tables 19 and 16, this shows that CRA transferability is not architecture-specific: ID CRA quantitatively predicts OOD CRA across a wide range of network depths and widths.

Connection to OOD performance prediction. Beyond confirming rank preservation, this 17-model set provides the data points used in the ID-CRA-vs-OOD-performance correlation analysis of Section 5.3: the (architecture, training-ε) combinations span a wide range of ID CRA values (from

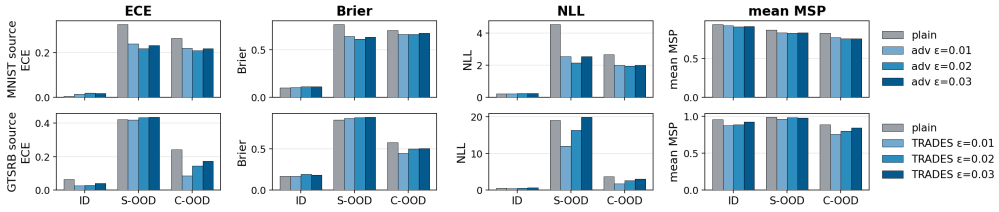


Figure 10: Full six-metric panel for MNIST-FC (top) and GTSRB-CNN (bottom): ECE, Brier, NLL and mean MSP per suite and model. Lower is better for ECE, Brier and NLL; mean MSP is a behavioral probe. Figure 4a in the main text shows ECE, NLL and mean MSP as trends over the training budget.

$\sim$46% on plain `mnist_fc` to $\sim$97% on adversarially trained wide/deep variants), which is necessary to populate the x-axis of Figure 4b. We therefore reuse the same set of models for two complementary purposes: demonstrating that rank ordering is architecture-agnostic, and demonstrating that ID CRA quantitatively predicts OOD performance metrics across that same architectural diversity.

E.2 Probabilistic OOD Performance: Tabular Form

Table 17 reports the numeric values plotted in Figure 4a of the main body: ECE, Brier, NLL, and mean maximum-softmax probability (mean MSP) for every (model, distribution) pair on both architectures. Lower is better for ECE, Brier, and NLL; mean MSP is a behavioral probe rather than a quality metric, with well-calibrated models expected to satisfy $\overline{\mathrm{MSP}}(\mathrm{ID}) > \overline{\mathrm{MSP}}(\mathrm{OOD})$. Adversarial training improves all three quality metrics on shifted data and restores the proper ID-vs-OOD confidence ordering, with returns saturating at moderate training-ε (MNIST adv $\varepsilon = 0.02$; GTSRB TRADES $\varepsilon = 0.01$). The bold values mark the best (lowest for ECE/Brier/NLL) per (architecture, distribution) cell.

E.3 Model-Selection Regret

A deployer who can verify only in-distribution chooses a model by its ID CRA. The *regret* of that choice on a shifted distribution is the OOD CRA of the OOD-best model minus the OOD CRA of the model chosen; zero means ID CRA picked the OOD winner. Table 18 reports it for every family and OOD axis, next to the regret of choosing by clean test accuracy and the expected regret of a uniformly random choice. Ties in ID CRA are listed for every tied candidate. The interval bootstraps the whole procedure: the shared ID and OOD property sets are resampled, the ID-best model is re-selected, and its regret re-evaluated, with ties resolved to the worst tied candidate.

Choosing by ID CRA recovers the OOD-best model outright in four of six settings and is within a tie of it in a fifth; in every setting it does at least as well as choosing by clean accuracy or at random. The exception is the seven-model CIFAR-10 family at $\varepsilon = 2/255$, where ID CRA selects Adv-1.5 (80% on CIFAR-10-C) while Adv-3 reaches 87%: a loss of 7 points, against 18 expected from a random choice, in the one family where several models are close. Two features of the design make the comparison easier than it would be in general: the four-model families contain one plain and three robust models, so almost any robustness signal selects a robust model; and the families are matched in clean accuracy by construction, so choosing by clean accuracy is close to arbitrary within them, which is why its regret on GTSRB is large. The random-choice column is the fair baseline.

F Checks: Verifiers and Smoothing

F.1 Multi-Verifier Sanity Check

Our main analysis uses α, β-CROWN (Wang et al., 2021) as the single verifier for every (model, distribution, ε) cell, since all comparisons in the body fix the verifier and property set so that the soundness gap $\rho_{f_\theta} - \hat{\rho}_{f_\theta}$ is held constant. To guard against the possibility that our rank-preservation finding is an artifact of a particular verifier's bound-tightness, we additionally re-run the MNIST-FC property suite through two complementary verifiers (Marabou (Wu et al., 2024) and NNV (Tran

Table 17: Probabilistic OOD performance per (model, distribution) pair on the matched-clean-accuracy MNIST-FC and GTSRB-CNN model families. ECE uses 15 equal-width bins on top-label confidence. Brier is the multiclass score $\frac{1}{N}\sum_i\sum_c(p_{i,c}-\mathbf{1}[y_i{=}c])^2$. NLL is per-sample cross-entropy. Mean MSP is the average top-class probability. **Bold** marks the best value in each (architecture, distribution) cell; ID rows are reported for reference.

Distribution	Model	ECE ↓	Brier ↓	NLL ↓	mean MSP
(a) MNIST-FC					
MNIST (ID)	Plain	**0.005**	**0.103**	**0.238**	0.934
	Adv-0.01	0.014	0.105	0.239	0.918
	Adv-0.02	0.020	0.113	0.258	0.907
	Adv-0.03	0.018	0.113	0.256	0.909
EMNIST (struct. OOD)	Plain	0.326	0.766	4.548	0.866
	Adv-0.01	0.239	0.643	2.554	0.828
	Adv-0.02	**0.218**	**0.611**	**2.159**	0.821
	Adv-0.03	0.232	0.632	2.555	0.827
PixMix-MNIST (corr. OOD)	Plain	0.264	0.703	2.682	0.824
	Adv-0.01	0.220	0.662	2.020	0.770
	Adv-0.02	**0.209**	**0.662**	**1.967**	0.752
	Adv-0.03	0.218	0.676	2.021	0.754
(b) GTSRB-CNN					
GTSRB (ID)	Plain	0.065	0.170	0.544	0.959
	TRADES-0.01	**0.028**	**0.169**	**0.447**	0.876
	TRADES-0.02	0.029	0.195	0.605	0.886
	TRADES-0.03	0.041	0.183	0.646	0.923
BTSD (struct. OOD)	Plain	0.422	**0.850**	18.985	0.988
	TRADES-0.01	**0.419**	0.867	**12.041**	0.963
	TRADES-0.02	0.434	0.877	16.364	0.982
	TRADES-0.03	0.434	0.881	19.857	0.980
PixMix-GTSRB (corr. OOD)	Plain	0.241	0.576	3.743	0.886
	TRADES-0.01	**0.085**	**0.446**	**1.765**	0.759
	TRADES-0.02	0.145	0.502	2.623	0.801
	TRADES-0.03	0.174	0.503	3.095	0.844

Table 18: Model-selection regret (points of OOD CRA). "Selected by ID CRA" lists the model(s) with the highest ID CRA and their OOD CRA; the regret columns give the OOD CRA lost relative to the OOD-best model when selecting by ID CRA (with a 95% interval from bootstrapping the selection procedure), by clean accuracy, and at random.

Family	OOD axis	ε	Selected by ID CRA (OOD CRA)	Regret	95% int.	Clean-acc. regret	Random regret
MNIST-FC	EMNIST	$2/255$	Adv-0.01/0.02/0.03 (55/54/53)	0/1/2	$[0, 6]$	0	1.5
MNIST-FC	PixMix	$2/255$	Adv-0.01/0.02/0.03 (68/70/68)	2/0/2	$[0, 8]$	2	4.0
GTSRB-CNN	BTSD	$2/255$	TRADES-0.02 (85)	0	$[0, 7]$	68	20.3
GTSRB-CNN	PixMix	$2/255$	TRADES-0.02 (37)	0	$[0, 3]$	35	10.2
CIFAR-CNN	CIFAR-10-C	$1/255$	Adv-1.5/3 (93/94)	1/0	$[0, 10]$	3	7.1
CIFAR-CNN	CIFAR-10-C	$2/255$	Adv-1.5 (80)	7	$[0, 14]$	19	17.7
CIFAR-CNN (TRADES)	CIFAR-10-C	$1/255$	TRADES-2/255/β6 (95)	0	$[0, 11]$	1	4.9
CIFAR-CNN (TRADES)	CIFAR-10-C	$2/255$	TRADES-0.01/β1 (79)	0	$[0, 18]$	0	11.4

et al., 2020; Lopez et al., 2023)) under the same 60-second per-instance timeout. The Marabou and NNV rows of Table 19 report the resulting CRA values.

Across all twelve (training-variant, distribution) cells of the MNIST-FC table, the rank ordering Plain $<$ Adv-0.01 $\approx$ Adv-0.02 $\approx$ Adv-0.03 is preserved by every verifier. Marabou, a complete verifier, decides every instance within the timeout and reproduces the α, β-CROWN CRA exactly on every cell, as two complete verifiers must on identical properties; NNV agrees with α, β-CROWN on all but one cell. Crucially, neither alternative verifier inverts the ordering at any ε on any distribution. The

Table 19: CRA (%) for MNIST-FC models across three test distributions and three verifiers. Columns under each distribution report CRA at $\varepsilon \in \{1, 2, 3, 4\}/255$ and the mean (*Avg*). The standard error on a single per-ε cell is $\leq$ 5pp ($\sqrt{p(1-p)/100}$); per-cell differences within $\pm$5pp should not be over-interpreted, and our claims are about rank order rather than absolute magnitudes. **Bold** marks the best model per column within each verifier block.

		MNIST (in-dist.)					**EMNIST** (struct. OOD)					**PixMix** (corrupt. OOD)				
Verifier	Model	$^1/_{255}$	$^2/_{255}$	$^3/_{255}$	$^4/_{255}$	Avg	$^1/_{255}$	$^2/_{255}$	$^3/_{255}$	$^4/_{255}$	Avg	$^1/_{255}$	$^2/_{255}$	$^3/_{255}$	$^4/_{255}$	Avg
α, β-CROWN	Plain	92	89	84	78	85.75	54	52	42	38	46.50	63	58	54	47	55.50
	Adv-0.01	**95**	**93**	92	90	92.50	55	**55**	**54**	**53**	**54.25**	**73**	68	66	64	67.75
	Adv-0.02	94	**93**	**93**	**92**	**93.00**	**57**	54	53	52	54.00	70	**70**	**68**	**67**	**68.75**
	Adv-0.03	93	**93**	**93**	**92**	92.75	55	53	53	51	53.00	70	68	**68**	65	67.75
MARABOU	Plain	92	89	84	78	85.75	54	52	42	38	46.50	63	58	54	47	55.50
	Adv-0.01	**95**	**93**	92	90	92.50	55	**55**	**54**	**53**	**54.25**	**73**	68	66	64	67.75
	Adv-0.02	94	**93**	**93**	**92**	**93.00**	**57**	54	53	52	54.00	70	**70**	**68**	**67**	**68.75**
	Adv-0.03	93	**93**	**93**	**92**	92.75	55	53	53	51	53.00	70	68	**68**	65	67.75
NNV	Plain	92	89	84	78	85.75	54	52	42	38	46.50	63	58	54	47	55.50
	Adv-0.01	**95**	**93**	92	90	92.50	55	**55**	**54**	**43**	51.75	**73**	68	66	64	67.75
	Adv-0.02	94	**93**	**93**	**92**	**93.00**	**57**	54	53	**52**	**54.00**	70	**70**	**68**	**67**	**68.75**
	Adv-0.03	93	**93**	**93**	**92**	92.75	55	53	53	51	53.00	70	68	**68**	65	67.75

rank-preservation observation reported in Section 5.1 therefore reflects a property of the underlying models rather than of the verification backend used to measure it.

F.2 RANDOMIZED SMOOTHING

To test the ℓ_2 instance of the framework, we apply randomized smoothing (Cohen et al., 2019) to the same base models, with $\sigma \in \{0.25, 0.5\}$, $N = 10^5$ noise samples, and failure probability $\alpha = 10^{-3}$ per input, on 100 points from each of the ID, structural-OOD and corruption-OOD sets, and report the certified ℓ_2 accuracy at radii $r \in \{0.25, 0.5\}$; abstentions count as not certified. The base models were not trained with Gaussian noise, so the smoothed Plain models abstain often, most visibly on GTSRB. Tables 20 and 21 give the results.

At every certified radius $r > 0$ and both σ, no model has lower smoothed certified accuracy than Plain, on either dataset (30 cells on MNIST-FC, 29 on GTSRB-CNN). Among the robust models the ID$\rightarrow$OOD ordering changes in several cells, more on GTSRB than on MNIST; the per-point certified radii were not retained, so the significance of those changes has not been assessed with a paired test, and we do not claim the fine ordering is preserved.

Table 20: Smoothed ℓ_2 certified accuracy (%) on MNIST-FC ($n = 100$ per set).

σ	Model	$r = 0.25$			$r = 0.5$		
		ID	structural	corruption	ID	structural	corruption
0.25	Plain	31	26	24	5	15	9
0.25	Adv-0.01	87	49	38	70	32	28
0.25	Adv-0.02	89	49	41	78	32	33
0.25	Adv-0.03	85	48	36	71	36	29
0.5	Plain	9	17	12	1	10	5
0.5	Adv-0.01	79	43	36	57	28	23
0.5	Adv-0.02	84	43	38	66	31	30
0.5	Adv-0.03	81	45	35	61	30	26

Table 21: Smoothed ℓ_2 certified accuracy (%) on GTSRB-CNN ($n = 100$ per set).

σ	Model	$r = 0.25$			$r = 0.5$		
		ID	structural	corruption	ID	structural	corruption
0.25	Plain	31	15	6	15	3	1
0.25	TRADES-0.01	55	41	14	42	32	11
0.25	TRADES-0.02	47	44	19	35	37	12
0.25	TRADES-0.03	47	35	18	37	24	8
0.5	Plain	10	1	1	4	0	0
0.5	TRADES-0.01	33	40	8	23	27	7
0.5	TRADES-0.02	23	40	9	18	35	6
0.5	TRADES-0.03	24	32	15	18	24	9