TEACHER–STUDENT DISAGREEMENT FOR UNCERTAINTY-AWARE HUMAN-IN-THE-LOOP AI

Anonymous authors
Paper under double-blind review

ABSTRACT

Large-scale classifiers are vulnerable to adversarial test-time attacks: small, deliberately crafted input perturbations that cause confident misclassification. Defenses such as adversarial training give only empirical robustness, and formal certification does not scale to modern models, so a deployed system still needs a mechanism for deciding when a prediction should be deferred to a human expert. Such *selective prediction* schemes typically use the model's own confidence, its maximum softmax probability, as the deferral trigger. However, our analysis shows that under attack this score is not merely unreliable but inverted: on CIFAR-10 (Krizhevsky et al., 2009) and MNIST (LeCun et al., 1998), attacked inputs receive *higher* confidence than clean ones, so a confidence-based rule defers less exactly when it should defer more. We propose a framework that pairs a standard-trained teacher model with a small ensemble of lightweight student models that are distilled from the teacher, adversarially trained, and uses teacher–student disagreement as a signal that the teacher may have been compromised. Against predictive entropy, mutual information and vote margin computed on the ensemble, and against nine selective-prediction and adversarial-detection baselines, disagreement is the strongest signal for deciding *which* teacher predictions to trust, that is, for ranking the teacher's wrong predictions ahead of its correct ones so that a human reviews the wrong ones. For instance, it ranks teacher errors with an area under the ROC curve (AUROC) of 87–98 where confidence scores 10–64. Adversarial detectors such as Mahalanobis distance detect attacks as well as disagreement does but cannot rank teacher errors. Because attacks crafted for the teacher rarely transfer to the robust students, the ensemble also serves as a fallback predictor (75–99% accuracy where the teacher is at 0–9%). The same mechanism sets the limits: on clean inputs confidence remains the better selector; measures computed on the students alone are blind to attacks that do not transfer to them; the students must be accurate enough to serve as a reference, a condition the practitioner can check from clean accuracy.

1 INTRODUCTION

Deep neural networks remain vulnerable to adversarial test-time attacks, in which subtle, human-imperceptible perturbations added to an input cause incorrect predictions (Szegedy et al., 2013; Goodfellow et al., 2014). Defenses such as adversarial training (Madry et al., 2017) give only empirical robustness and fail against adaptive adversaries (Athalye et al., 2018); certified defenses (Wong & Kolter, 2018; Cohen et al., 2019; Gowal et al., 2018) give guarantees but scale poorly to large models. A deployed system in a safety-critical domain therefore still needs a mechanism for deciding, input by input, when its prediction can be trusted and when it should be deferred to a human expert. This is the *selective prediction* setting (Geifman & El-Yaniv, 2017).

Existing human-in-the-loop frameworks use the model's own confidence as the signal for intervention (Hendrycks & Gimpel, 2016; Geifman & El-Yaniv, 2017; Madras et al., 2018; Mozannar & Sontag, 2020). Under attack this is precisely the wrong signal, because attacks are designed to maximize confidence in the wrong class (Carlini & Wagner, 2017; Stutz et al., 2020). A preliminary experiment makes this concrete: we attack a standard-trained ResNet-50 teacher (94.8% clean on CIFAR-10) with PGD (Madry et al., 2017), the standard iterative attack, at $\epsilon = 8/255$, a perturbation invisible to a human. The teacher drops to 0% accuracy while its mean softmax confidence *rises* to 1.000, above its clean value (Section 4.3). Confidence does not merely weaken under attack; it points

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.

the wrong way, consistent with prior reports that adversarial examples are confidently wrong (Carlini & Wagner, 2017; Stutz et al., 2020; Goodfellow et al., 2014).

We argue that comparing the teacher against a small ensemble of adversarially robust *student models* gives a far better deferral signal, and three design choices are deliberate. We add a separate model rather than adversarially train the teacher, since adversarial training lowers clean accuracy and is often costly or unavailable on a large or third-party teacher, which we leave untouched. We adversarially train the students so they keep predicting correctly under the attack that breaks the teacher, making their disagreement informative rather than noise. And we distill from the teacher so the students track its clean decision boundary and stay accurate enough to be a reference, so a disagreement signals an attack, not a merely weaker student. The students are also *diverse* in depth, kernel sizes, downsampling, channel widths and the absence of residual and normalization layers, so a perturbation tuned to the teacher rarely fools them the same way. At inference the teacher predicts, the students are the reference, and we defer on disagreement. We also evaluate three finer-grained ensemble measures: predictive entropy of the averaged distribution (total uncertainty), mutual information (the epistemic part, high when confident members disagree (Houlsby et al., 2011; Smith & Gal, 2018)), and the vote margin between the two most-voted classes.

Our contributions are:

1. A hybrid selective-prediction framework in which a standard-trained teacher is paired with a distilled, adversarially trained student ensemble whose disagreement with the teacher drives deferral, decoupling it from the teacher's own confidence. We instantiate seven deferral rules and evaluate them threshold-free against nine baselines from the selective-prediction and adversarial-detection literatures. Disagreement is the only signal that does both jobs: every teacher-derived score inverts under attack, and the detectors flag the attack but cannot rank the teacher's errors.
2. The mechanism reliably flags adversarial manipulation of the teacher: disagreement separates attacked from clean inputs with detection AUROC 83–99.5 across CIFAR-10, GTSRB and MNIST under FGSM, PGD, C&W and AutoAttack, matching the strongest detector (Mahalanobis) while, unlike it, also ranking which surviving predictions are wrong. On GTSRB under PGD, for instance, it ranks teacher errors at AUROC 96.4 against 10.1 for confidence, and at a 50% query budget recovers 58.8% oracle accuracy, the perfect-ranker value, where confidence recovers the 50.0% of random deferral.
3. We explain what bounds the method. Its clean false-alarm rate is set by the student–teacher accuracy gap, since disagreement fires whenever a student is wrong; this fixes the operating condition of students within roughly 20 points of the teacher, met on MNIST, GTSRB and CIFAR-10 (gaps 0.5, 4, 19) and failed on Tiny ImageNet (gap 44). These results are for ℓ_∞-attacked image classification with a hard-label oracle; other threat models, structured-output tasks, and other metrics are left open.

2 Related Work

Adversarial robustness. Adversarial vulnerability was identified by Szegedy et al. (2013); Goodfellow et al. (2014) introduced FGSM, a single-step attack that moves every pixel by ϵ in the direction of the loss gradient's sign, and Madry et al. (2017) introduced PGD, its iterative refinement, together with the adversarial-training scheme we use for our students. Carlini & Wagner (2017) proposed the C&W attack, which maximizes the margin between the wrong and correct class and thereby the model's confidence in the wrong answer, and Croce & Hein (2020) AutoAttack, a parameter-free ensemble of four attacks that remains the standard robustness benchmark; sound evaluation against adaptive adversaries continues to be emphasized as the correct methodology (Tramer et al., 2020). In our framework adversarial training is not a defense for the teacher, which stays standard-trained; it is how the students are made robust, so that they keep predicting correctly under attacks that break the teacher and their disagreement becomes informative.

Human-in-the-loop learning. Learning to defer (Madras et al., 2018; Mozannar & Sontag, 2020), human–machine complementarity (Wilder et al., 2020) and differentiable triage (Okati et al., 2021) all rely on model confidence as the deferral criterion, which attacks inflate (Carlini & Wagner, 2017; Stutz et al., 2020), and the difficulty of calibrating this criterion has kept it an active concern (Verma & Nalisnick, 2022; Mosqueira-Rey et al., 2023). That the deferral mechanism is itself an attack

surface has been made explicit only recently: Montreuil et al. (2025) show that adversarial perturbations can manipulate the allocation policy of a learning-to-defer system, misrouting queries or overloading the human, and give the first robustness guarantees for that setting. Our work addresses the complementary failure: not the routing rule but the confidence signal it is built on, which under attack does not merely miscalibrate but inverts. We replace it with teacher–student comparison, and include a logistic-regression selector fit on clean data as a baseline; it is best on clean data but under attack fails exactly as raw confidence does, because it has learned to trust the score that attacks inflate.

Uncertainty, selective prediction and adversarial detection. MC dropout (Gal & Ghahramani, 2016) and deep ensembles (Lakshminarayanan et al., 2017) give practical uncertainty estimates; networks are nonetheless poorly calibrated (Guo et al., 2017) and fail under shift (Ovadia et al., 2019), and selective classification itself can behave unexpectedly under distribution shift, of which an adversarial perturbation is an adversarially chosen instance (Jones et al., 2020). Houlsby et al. (2011) introduced predictive entropy and mutual information as epistemic-uncertainty measures, which Smith & Gal (2018) applied to adversarial detection. We adopt the threshold-free selective-classification metrics of Geifman & El-Yaniv (2017; 2019). Our baselines are the single-model scores maximum softmax probability (Hendrycks & Gimpel, 2016), max-logit (Basart et al., 2022), energy (Liu et al., 2020) and ODIN (Liang et al., 2017), and the detectors Mahalanobis distance (Lee et al., 2018) and feature squeezing (Xu et al., 2017). Such detectors are not automatically robust: a line of recent work shows that out-of-distribution and adversarial detectors can themselves be evaded by adversaries aware of them (Azizmalayeri et al., 2022), which is why we evaluate our signal under an ensemble-aware adaptive attack. The two families answer different questions, "is the teacher wrong?" versus "was the input tampered with?"; we show that disagreement answers the first far better than any single-model score under attack and the second as well as the strongest detector.

Ensembles for robustness. Adversarial examples transfer imperfectly between architectures, which Tramèr et al. (2017) and Pang et al. (2019) exploit to build robust ensembles. Knowledge distillation (Hinton et al., 2015) trains a small student to match a large teacher; we follow adversarially robust distillation (Goldblum et al., 2020) in distilling from the teacher's clean output. Unlike prior work, our ensemble is not the classifier: the framework always returns the teacher's prediction, and the smaller, less accurate students serve as a robust reference to compare against. The signal is the *comparison*, not the ensemble's own uncertainty, which Section 4.3 shows is blind to attacks that target the teacher.

3 METHODOLOGY

A practitioner has a high-accuracy *teacher model* f_θ and needs to decide, input by input, whether to accept its prediction or have a human check it, given that the teacher may be attacked and its confidence is then unreliable. We train a small ensemble of lightweight *student models* that learn the task from the teacher but are adversarially robust and built differently, so an attack on the teacher is unlikely to fool them the same way. At inference the teacher predicts and the students act as a reference; from their agreement or disagreement we derive a family of uncertainty measures and defer whenever the chosen measure distrusts the teacher. The framework thus has four parts, detailed below: training the teacher (Section 3.1), training the student ensemble (Section 3.2), computing ensemble uncertainty measures (Section 3.3), and turning them into a deferral decision (Section 3.4). Two facts make this work: attacks maximize the teacher's confidence in the wrong class (Carlini & Wagner, 2017), so its softmax is untrustworthy exactly when it errs, and perturbations transfer imperfectly across architectures (Tramèr et al., 2017; Pang et al., 2019), so an attacked teacher disagrees with unattacked students. Because our attacks target the teacher, they usually do not reach the students (Table 3); a measure on the students alone therefore sees nothing, and only the comparison registers the attack.

3.1 TEACHER MODEL TRAINING

We train a high-capacity teacher model $f_\theta : \mathcal{X} \to \Delta^{K-1}$ to maximize clean accuracy on each benchmark, using ResNet-50 (He et al., 2016) with a CIFAR-style stem as the backbone for MNIST, CIFAR-10, and GTSRB, and ConvNeXt-Base (Liu et al., 2022) for Tiny ImageNet, where the larger label space and higher input resolution benefit from greater representational capacity. In all cases the

teacher is trained with standard supervised cross-entropy:

$$\mathcal{L}_{\text{teacher}}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[-\log f_\theta(x)_y\big], \tag{1}$$

where $f_\theta(x)_y$ denotes the predicted probability for the correct class y. The teacher is deliberately *not* adversarially trained: it represents the high-accuracy model a practitioner already has, and adversarially retraining a large model is often costly or unavailable (e.g. a third-party model). Our design confines adversarial training to the students, one to two orders of magnitude smaller than the teacher.

3.2 Student Ensemble: Adversarial Training and Distillation

We train an ensemble of M student models $\{f_{\phi_j} : \mathcal{X} \to \Delta^{K-1}\}_{j=1}^{M}$ that are diverse in architecture, lightweight, and *adversarially trained* so that the ensemble remains informative under adversarial perturbation. In our experiments we set $M = 6$: enough for a stable majority vote and for an attack to fail to transfer to some members, while the whole ensemble stays at a fraction of the teacher's parameters (6.6M against 23M–89M on CIFAR-10).

Architectural Diversity. To maximize disagreement on adversarial examples (due to limited attack transferability), the six student models are lightweight (0.65–1.43M parameters each), unnormalized convolutional networks with systematically varied structural choices. We denote the student models $S_1, \ldots, S_6$. All six use ReLU activations, end with adaptive average pooling to a 4×4 spatial grid, and feed a two-layer fully-connected classifier head ($C \cdot 16 \to 256 \to K$); differences are concentrated in the convolutional trunk:

1. **Depth and block structure**: S_1 and S_2 use 5 plain 3×3 convolutions; S_3 uses a shallower 4-layer trunk; S_4 uses 7 convolutions in two bottleneck blocks ($3\times3 \to 1\times1$ compress $\to 3\times3$); S_5 uses a 6-layer trunk in which every second convolution is dilated (dilation 2), enlarging the receptive field without extra depth; and S_6 uses the shallowest trunk of all, 3 convolutions.
2. **Downsampling**: MaxPool (S_1, S_3, S_4), strided 3×3 convolutions (S_2, S_6), or AvgPool (S_5).
3. **Kernel sizes**: 3×3 throughout (S_1, S_2, S_3); mixed 3×3 and 1×1 (S_4); 3×3 with alternating dilation (S_5); or a 5×5 stem followed by 3×3 (S_6).
4. **Channel widths**: $3 \to 32 \to 64 \to 128$ for S_1 and S_2; $3 \to 64 \to 128$ for S_3; the bottleneck progression $3 \to 32 \to 16 \to 32 \to 64 \to 32 \to 64 \to 128$ for S_4; the widest, $3 \to 48 \to 96 \to 192$, for S_5; and $3 \to 64 \to 128 \to 128$ for S_6.

No student uses normalization: this adds diversity from the normalized-residual teacher and avoids batch-norm statistics being shared across clean and adversarial inputs. The students are then deep and unnormalized, so they use He initialization (He et al., 2015); under the PyTorch default the deepest start as near-constant functions and never train on the 43- and 200-class benchmarks.

Adversarial Training and Knowledge Distillation. Each student model is trained with a hybrid objective that combines knowledge distillation (Hinton et al., 2015) from the teacher with PGD-based adversarial training (Madry et al., 2017). For an input x with ground-truth label y, let z_s and z_t denote the student and teacher logits. The per-input distillation loss applied at temperature T is

$$\mathcal{L}_{\text{KD}}(x, y) = \alpha \cdot T^2 \cdot \text{KL}\big(\text{softmax}\tfrac{z_s}{T} \,\big\|\, \text{softmax}\tfrac{z_t}{T}\big) + (1-\alpha)\cdot \text{CE}(z_s, y), \tag{2}$$

with temperature $T = 4.0$ and distillation weight $\alpha = 0.7$; the T^2 scaling follows standard practice (Hinton et al., 2015). We apply this loss on both clean inputs x and a PGD-perturbed counterpart $x' = x + \delta^\star$, and combine the two:

$$\mathcal{L}(x, y) = 0.2 \cdot \mathcal{L}_{\text{KD}}(x, y) + 0.8 \cdot \mathcal{L}_{\text{KD}}(x', y), \tag{3}$$

where, following adversarially robust distillation (Goldblum et al., 2020), the teacher logits z_t in the second term are computed on the *clean* input x, so that a student model is never distilled from a teacher that has itself been fooled by transfer. The weights 0.2 and 0.8 are fixed, with the adversarial term dominant so the students stay robust. The PGD perturbation $\delta^\star$ is generated on the student's

own loss surface within an ℓ_∞ ball of radius ϵ in $[0, 1]$ pixel space ($8/255$ for CIFAR-10, GTSRB and Tiny ImageNet; 0.3 for MNIST), using 4 steps of size $\epsilon/4$ from a random start, with ϵ ramped linearly from 0 over the first 10 epochs. Per-dataset input normalization is folded into each model, so the budget is identical in pixel space for every model and every attack.

All students are trained for 200 epochs with SGD (momentum 0.9, weight decay 5×10^{-4}) at an initial learning rate of 0.02 (0.005 on MNIST and Tiny ImageNet), decayed by 0.1 at epochs $\{80, 140, 180\}$. We use a batch size of 256, mixed-precision (AMP) training, and gradient-norm clipping at 1.0. We select the checkpoint with the highest 10-step-PGD robust accuracy on a held-out $2{,}000$-image training slice, reporting on the test set only. The resulting student models reach 68–76% clean accuracy and 31–34% robust accuracy on CIFAR-10, where *robust accuracy* is accuracy under a white-box 20-step PGD attack on the student itself (white-box: the attacker has the model's weights and gradients). On GTSRB they reach $\approx 91\%$ clean / 48% robust, and on MNIST 98.5–99.4% / 80–91%. On Tiny ImageNet, a 200-class benchmark at 64×64 resolution, the same architectures reach only 20–26% clean and 7–10% robust accuracy: networks with 0.65M parameters that were designed for 32×32 inputs do not have the capacity to learn a task of that size under adversarial training, and the student–teacher accuracy gap widens to 44 points. These Tiny ImageNet students are below the capacity the method requires; with student–teacher gaps of 0.5, 4, 19 and 44 points, the four benchmarks bracket where disagreement stops being informative (Section 5).

3.3 Ensemble Uncertainty Measures

For an input x, let M be the number of students, K the number of classes, f_{ϕ_j} the j-th student and f_θ the teacher. We write $p_j = f_{\phi_j}(x) \in \Delta^{K-1}$ for the j-th student's class distribution (a vector over the K classes), $(p_j)_k$ for its k-th entry, $\hat{y}_j = \arg\max_k (p_j)_k$ for its hard prediction, $\hat{y}_T = \arg\max_k f_\theta(x)_k$ for the teacher's prediction and $c_T(x) = \max_k f_\theta(x)_k$ for the teacher's confidence (its top softmax probability). We define four signals. *Teacher–student disagreement*: with $\hat{y}_E$ the students' majority vote and $\mathbb{1}[\cdot]$ the indicator, $D(x) = \mathbb{1}[\hat{y}_T \neq \hat{y}_E]$, which is 1 when teacher and ensemble disagree and 0 otherwise. *Predictive entropy*: the Shannon entropy $\mathcal{H}[\bar{p}(x)]$ of the mean distribution $\bar{p}(x) = \frac{1}{M}\sum_j p_j$, the ensemble's total uncertainty (aleatoric, from ambiguity in the input, plus epistemic, from the models' lack of knowledge). *Mutual information*: following Houlsby et al. (2011); Smith & Gal (2018), the epistemic part $\mathcal{I}(x) = \mathcal{H}[\bar{p}(x)] - \frac{1}{M}\sum_j \mathcal{H}[p_j]$, the total uncertainty minus the mean of the members' own entropies, which is high when members are individually confident but disagree. *Vote margin*: with $v_k(x)$ the fraction of students voting for class k and $v_{(1)}, v_{(2)}$ its two largest values, $m(x) = v_{(1)}(x) - v_{(2)}(x) \in [0, 1]$, the gap between the two most-voted classes.

3.4 Deferral Rules and Baselines

A deferral rule is a test on these quantities: if it passes, the input goes to the human oracle (a reviewer assumed to answer correctly; see Section 5); otherwise the framework returns the teacher's prediction $\hat{y}_T$. The students never predict on the framework's behalf; they only decide whether the teacher should be checked. We compare seven rules: **S1** (low confidence): defer if $c_T(x) < \tau$, the standard baseline. **S2** (disagreement): defer if $D(x) = 1$, i.e. whenever the teacher's class differs from the student majority. **S3** (S1 *and* S2): defer if $D(x) = 1$ and $c_T(x) < \tau$, the conservative joint rule. **S4** (S1 *or* S2): defer if $D(x) = 1$ or $c_T(x) < \tau$, the permissive joint rule. **S5** (entropy): defer if $\mathcal{H}[\bar{p}(x)] > \tau_H$. **S6** (vote margin): defer if $m(x) < \tau_V$. **S7** (mutual information): defer if $\mathcal{I}(x) > \tau_I$. The fixed-threshold results of Appendix E use $\tau = \tau_H = \tau_V = 0.5$, $\tau_I = 0.3$. The threshold-free metrics below do not use these values: they rank all test inputs from "defer first" to "defer last" by the quantity each rule thresholds ($1 - c_T$, D, $D\,(1 - c_T)$, $\max(D, 1 - c_T)$, $\mathcal{H}[\bar{p}]$, $1 - m$, $\mathcal{I}$) and evaluate the whole ranking. S2 yields only two ranks; S3 and S4 refine them by confidence.

Baselines. Nine scores computed on the same teacher: max-logit (Basart et al., 2022); teacher entropy; energy (Liu et al., 2020); ODIN (Liang et al., 2017) (temperature 1000, input perturbation 0.0014); Mahalanobis distance (Lee et al., 2018) from the penultimate features to the nearest class mean, with means and shared covariance fit on $5{,}000$ clean training images; MC dropout (Gal & Ghahramani, 2016) on penultimate features ($p = 0.2$, 20 samples); feature squeezing (Xu et al., 2017) (5-bit depth reduction and 3×3 median filter); a *deep ensemble* baseline (Lakshminarayanan et al., 2017) in which the students predict on their own with entropy as the score; and a *learned deferral*

selector, logistic regression on the teacher and ensemble scores fit to predict teacher errors on a clean 20% split of the test set, with all selective-prediction results reported on the remaining 80%.

3.5 Datasets and Metrics

We evaluate on MNIST (LeCun et al., 1998) (10 classes), CIFAR-10 (Krizhevsky et al., 2009) (10), GTSRB (Stallkamp et al., 2012) (43 traffic-sign classes, imbalanced, real driving conditions) and Tiny ImageNet (Le et al., 2015) (200 classes, 64×64); the first three are used at 32×32 (details in Appendix C).

A fixed threshold can make an uninformative rule look perfect by deferring everything, so we report threshold-free metrics and relegate fixed-threshold numbers to Appendix E. *(1) Selective risk.* Ranking inputs by a score and letting the teacher answer the most trusted fraction c (the coverage) gives a risk–coverage curve whose area (AURC, ↓) (Geifman & El-Yaniv, 2019) summarizes selective prediction across all thresholds. We also report the *error-detection AUROC* (↑): the probability that a random wrong teacher prediction receives a higher deferral score than a random right one (50 is chance; below 50 means the score prefers the wrong predictions). *(2) Oracle accuracy at a matched human-query budget.* The budget $b \in [0, 1]$ is the fraction of the N test inputs the framework may send to the oracle. We rank inputs by the rule, send the top $b\,N$ to the oracle, who answers correctly, and let the teacher answer the rest, giving $\mathrm{Acc}_{\text{oracle}}(b) = \big(\mathrm{Correct}_{\text{teacher}}(\text{non-deferred}) + b\,N\big)/N$. A perfect ranker defers only errors and reaches $\mathrm{acc}_T + b$; a random one wastes a fraction acc_T of its queries on inputs the teacher had right and reaches $b + (1-b)\,\mathrm{acc}_T$ (e.g. 58.8% vs. 54.4% at $b = 50\%$ when $\mathrm{acc}_T = 8.8\%$). Matching b is essential, since a rule that defers more would otherwise look better simply because the human answers more. *(3) Adversarial-detection AUROC* (↑): separation of attacked from clean inputs, the only informative metric when an attack drives the teacher to 0%. *(4) Fallback accuracy*: the student majority's accuracy alongside the teacher's, since the ensemble can serve as a fallback predictor.

4 Experiments and Results

4.1 Setup

For each dataset, we evaluate every strategy and baseline on the clean test set and on adversarial test sets crafted *against the teacher model* with FGSM (Goodfellow et al., 2014), PGD (Madry et al., 2017) (20 steps; 40 steps of size 0.01 on MNIST), a C&W-loss attack (Carlini & Wagner, 2017) (margin loss, 100 PGD steps, ℓ_∞-projected), and AutoAttack (Croce & Hein, 2020) (standard version), all at $\epsilon = 8/255$ (0.3 on MNIST) in pixel space. We also run an *ensemble-aware adaptive* attack in which the adversary knows the students: 20-step PGD on the joint objective $\mathrm{CE}(f_\theta(x'), y) + \sum_j \mathrm{CE}(f_{\phi_j}(x'), y) - \lambda \sum_j \mathrm{KL}\big(f_\theta(x') \,\|\, f_{\phi_j}(x')\big)$ with $\lambda = 1$, which fools all models while pulling the students toward the teacher's output to suppress disagreement. The main text uses $\lambda = 1$; Appendix H sweeps $\lambda \in \{0.3, 1, 3, 10\}$ on GTSRB.

Metrics. Table 1 (error-detection AUROC) asks whether a rule ranks the teacher's wrong predictions ahead of its right ones; this is the selective-prediction question, the one that matters for deferral. Table 4 in Appendix A (adversarial-detection AUROC) asks whether a score reacts to an input being attacked, regardless of whether the teacher survives. A score can excel at one and fail the other, and disagreement is the only one strong on both (Section 4.3). Error ranking and oracle accuracy are defined only where the teacher keeps some correct predictions: GTSRB, where the teacher is partially broken under every attack (6–27%), is therefore the most informative benchmark, while on CIFAR-10 and MNIST every multi-step attack drives the teacher to 0% at the standard budget, so Table 1 shows those attacks at reduced budgets ($2/255$, 0.1) taken from the full sweeps of Appendix G. Oracle accuracy at matched budgets and the full teacher-versus-ensemble accuracies are in Tables 2 and 3 (Appendix A); the header rows of Table 1 carry the accuracies quoted in the text.

4.2 Clean Data Performance

On clean data the teacher's confidence is a strong selector and the ensemble rules do not beat it: on CIFAR-10, S1 ranks teacher errors at AUROC 92.9 (Table 1, Clean column) with AURC 0.64

Table 1: Error-detection AUROC (↑): how well each score ranks the teacher's wrong predictions ahead of its right ones. Multi-step attacks on CIFAR-10, MNIST and Tiny ImageNet are shown at reduced budgets ($2/255$, 0.1, $2/255$) where the teacher is partially broken (full sweeps in Appendix G); at full budget only attack detection (Table 4) is defined. The two header rows give the teacher's and the student majority's accuracy under each condition; the gap between them is the fallback available when the framework defers. AA = AutoAttack; Adapt = ensemble-aware adaptive attack. Deep Ensemble predicts with the student majority, so its row ranks *ensemble* errors. Best per column in **bold**; our strategies above the rule.

	GTSRB ($8/255$)						CIFAR-10 ($8/255$)		CIFAR-10 ($2/255$)			MNIST (0.3)		MNIST (0.1)			Tiny ($8/255$)		Tiny ($2/255$)		
	Clean	FGSM	PGD	C&W	AA	Adapt	Clean	FGSM	PGD	C&W	AA	Clean	FGSM	PGD	C&W	AA	Clean	FGSM	PGD	C&W	AA
Teacher acc. (%)	94.7	27.3	8.8	9.2	6.5	15.3	94.8	37.4	7.9	5.7	3.4	99.6	9.9	25.5	22.1	2.4	71.0	9.7	8.2	8.6	3.5
Students acc. (%)	90.7	84.2	85.2	84.4	86.1	60.7	76.0	74.8	75.7	75.7	75.8	99.1	97.8	98.8	98.8	98.9	26.6	25.3	26.4	26.4	26.4
S1 Low Confidence	96.9	63.8	10.1	11.0	46.7	20.8	**92.9**	58.8	6.1	4.8	39.8	90.1	78.3	32.6	29.7	74.5	86.1	44.2	18.1	20.6	75.8
S2 Disagreement	89.4	**96.7**	**96.4**	**95.6**	**97.6**	83.7	69.2	87.1	**93.4**	**93.2**	94.7	64.9	98.9	99.4	99.4	99.6	61.0	**76.2**	78.3	76.6	81.2
S3 S1 and S2	91.9	96.7	**96.4**	95.4	**97.6**	83.6	74.2	**88.1**	92.5	92.4	94.7	64.9	98.6	99.4	99.4	99.6	80.6	73.6	67.6	65.2	88.4
S4 S1 or S2	95.1	96.3	92.8	91.2	96.5	68.9	82.7	86.7	88.0	87.7	93.2	89.9	**98.9**	98.8	98.8	99.6	62.5	75.2	76.3	74.9	82.8
S5 Predictive Entropy	91.6	90.0	88.2	88.4	90.5	**88.0**	72.8	65.6	82.6	84.6	88.7	98.2	50.9	64.5	61.5	65.1	59.1	68.6	73.3	72.5	79.0
S6 Vote Margin	71.2	62.0	58.6	59.2	58.5	62.8	61.3	56.5	62.0	62.1	62.9	72.0	50.4	51.1	51.1	50.9	55.8	61.9	64.3	62.0	65.0
S7 Mutual Information	86.6	88.9	87.0	87.2	89.1	84.1	69.0	62.9	77.3	80.2	84.8	93.8	26.9	57.7	51.6	28.1	41.3	32.2	26.8	27.0	19.9
Max-Logit	95.8	58.4	9.4	11.7	40.6	19.2	90.4	55.5	5.2	4.8	35.0	89.4	77.1	29.1	27.5	70.3	83.4	34.5	12.1	18.8	64.9
Teacher Entropy	96.7	62.2	9.9	11.3	46.0	20.5	92.8	58.7	6.0	4.8	39.5	90.0	77.9	31.8	29.3	73.3	85.0	37.1	15.2	21.0	73.1
Energy	95.5	57.1	9.4	11.9	40.6	19.0	90.2	55.3	5.2	4.8	35.0	89.3	76.8	28.5	27.3	69.8	80.5	26.7	11.0	19.8	62.4
ODIN	89.9	53.7	9.4	11.6	38.2	18.5	74.8	48.8	4.7	4.6	25.6	89.7	76.6	27.8	26.8	64.0	76.9	23.6	10.6	18.6	52.7
Mahalanobis	96.8	72.0	47.1	57.1	95.4	49.6	91.6	62.3	28.2	32.5	91.9	95.7	70.6	71.1	67.3	**100.0**	82.5	68.6	69.8	70.2	**97.1**
MC Dropout	95.8	59.9	11.4	13.3	47.9	22.6	92.7	58.4	5.9	4.8	37.9	89.4	77.8	32.0	29.5	73.3	84.9	37.1	15.3	21.1	73.5
Feature Squeezing	94.7	69.9	27.6	29.4	85.0	40.1	80.6	76.7	89.6	89.6	**99.0**	94.3	79.0	64.4	67.0	99.9	71.2	75.9	**89.2**	**87.7**	95.8
Learned Deferral	**97.8**	88.3	25.7	31.1	74.5	42.2	91.1	57.2	5.5	4.9	37.5	**98.8**	94.4	68.3	66.9	99.9	**86.4**	53.1	51.0	47.0	94.3
Deep Ensemble (students predict)	94.2	91.8	90.9	90.5	91.6	82.5	83.1	82.8	82.9	82.9	82.9	98.2	96.9	97.8	97.8	97.9	74.5	74.6	74.6	74.4	74.6

Table 2: Oracle accuracy (%, ↑) at matched human-query budgets of 25% and 50%, with the fraction of deferred inputs the teacher had already answered correctly (*wasted slots*, %, ↓) in parentheses. A perfect ranker defers only errors and reaches $\mathrm{acc}_T + b$ with no waste; random deferral reaches $b + (1-b)\,\mathrm{acc}_T$, wasting slots at the teacher's accuracy rate. Wasted slots are derived as $(\mathrm{acc}_T + b - \mathrm{oracle})/b$. Best oracle accuracy per column in **bold**. Conditions are those where the teacher retains non-zero accuracy.

Teacher acc. (%)	CIFAR-10 FGSM 37.4		GTSRB FGSM 27.3		GTSRB PGD 8.8		GTSRB Adapt 15.3		Tiny FGSM 9.7	
	@25%	@50%	@25%	@50%	@25%	@50%	@25%	@50%	@25%	@50%
Perfect ranker	62.4 (0)	87.4 (0)	52.3 (0)	77.3 (0)	33.8 (0)	58.8 (0)	40.3 (0)	65.3 (0)	34.7 (0)	59.7 (0)
Random deferral	53.0 (37)	68.7 (37)	45.5 (27)	63.6 (27)	31.6 (9)	54.4 (9)	36.5 (15)	57.6 (15)	32.3 (10)	54.8 (10)
S1 Low Confidence	55.9 (26)	72.2 (30)	48.6 (15)	68.4 (18)	26.1 (31)	50.0 (18)	32.2 (32)	51.0 (29)	32.7 (8)	54.3 (11)
S2 Disagreement	60.2 (9)	83.2 (8)	**52.3** (0)	**77.3** (0)	**33.8** (0)	**58.8** (0)	40.2 (0)	65.2 (0)	33.7 (4)	57.6 (4)
S3 S1 and S2	**60.9** (6)	**83.4** (8)	52.2 (0)	77.2 (0)	**33.8** (0)	**58.8** (0)	40.2 (0)	65.2 (0)	33.6 (4)	57.0 (5)
S4 S1 or S2	60.1 (9)	83.2 (8)	**52.3** (0)	77.2 (0)	**33.8** (0)	**58.8** (0)	**40.2** (0)	**65.2** (0)	33.5 (5)	57.6 (4)
S5 Predictive Entropy	56.6 (23)	73.7 (27)	52.1 (1)	75.8 (3)	**33.8** (0)	58.7 (0)	40.2 (0)	64.5 (2)	33.5 (5)	56.8 (6)
S6 Vote Margin	55.8 (26)	70.7 (33)	49.8 (10)	66.9 (21)	32.8 (4)	55.3 (7)	39.7 (2)	59.8 (11)	33.2 (6)	56.5 (6)
S7 Mutual Information	55.5 (27)	73.4 (28)	51.7 (2)	75.3 (4)	**33.8** (0)	58.6 (0)	39.5 (3)	63.8 (3)	30.0 (19)	52.9 (14)
Mahalanobis	55.7 (27)	72.9 (29)	49.6 (11)	70.5 (14)	32.1 (7)	53.9 (10)	37.2 (12)	57.3 (16)	34.1 (2)	57.4 (5)
Feature Squeezing	59.7 (11)	78.6 (18)	49.4 (12)	69.6 (15)	30.4 (13)	50.4 (17)	36.8 (14)	55.1 (20)	**34.4** (1)	**58.3** (3)

(Table 7), S2 at 69.2 / 3.12, and the best ensemble rule S4 at 82.7 / 1.17; GTSRB shows the same ordering (S1 96.9, S4 95.1, S2 89.4). The header rows explain why: the students are 19 points behind the teacher on CIFAR-10 and 4 on GTSRB, and a student that is wrong disagrees with a teacher that is right, so disagreement raises false alarms on 23%, 7% and 0.5% of the inputs the teacher gets right on CIFAR-10, GTSRB and MNIST. MNIST, with students at 99.1%, is the exception: ensemble entropy (S5, 98.2) and mutual information (S7, 93.8) beat confidence (90.1). We claim no clean-data advantage; the clean cost of disagreement is a false-alarm rate set by student capacity.

4.3 Performance Under Adversarial Attacks

Confidence is inverted. Under every multi-step attack the confidence family (S1, max-logit, entropy, energy, ODIN, MC dropout) separates attacked from clean inputs with AUROC 0–12 on CIFAR-10 and MNIST (Table 4). These are not chance values but inversion: the CIFAR-10 teacher's mean confidence rises from 0.983 (clean) to 1.000 (PGD) as accuracy falls to 0%, so any rule built on confidence defers *less* under attack. On GTSRB it ranks teacher errors at AUROC 10 under PGD:

Table 3: Fallback accuracy (%): the teacher versus the student majority vote under each condition. Attacks are crafted against the teacher; the students are reached only by transfer (or, for Adapt, attacked directly).

Dataset	Model	Clean	FGSM	PGD	C&W	AA	Adapt
CIFAR-10	Teacher	94.8	37.4	0.0	0.0	0.0	0.0
	Students	76.0	74.8	75.3	75.3	75.3	55.8
GTSRB	Teacher	94.7	27.3	8.8	9.2	6.5	15.3
	Students	90.7	84.2	85.2	84.4	86.1	60.7
MNIST	Teacher	99.6	9.9	0.0	0.0	0.0	0.0
	Students	99.1	97.8	98.9	98.8	98.8	87.2
Tiny ImageNet	Teacher	71.0	9.7	0.0	0.0	0.0	0.3
	Students	26.6	25.3	25.7	25.7	25.6	16.8

it prefers the wrong predictions. The learned selector, fit on clean data, inherits the failure (0.1 on CIFAR-10 PGD; 25.7 on GTSRB).

Disagreement is the best selective-prediction signal. On GTSRB, S2 ranks teacher errors at AUROC 96.7 (FGSM), 96.4 (PGD), 95.6 (C&W) and 97.6 (AutoAttack), against 10–64 for confidence and 47–72 for Mahalanobis; with a 50% budget under PGD it recovers 58.8% oracle accuracy, the perfect-ranker value, where confidence recovers 50.0%, the random value (Table 2). On CIFAR-10 under FGSM, S2/S3 reach AUROC 87–88 and AURC 36 against 59 / 59.4 for confidence and beat every baseline including feature squeezing (76.7 / 44.7); at $2/255$, where the teacher survives PGD at 7.9%, S2 ranks errors at 93–95 under PGD, C&W and AutoAttack against 5–40 for confidence. On MNIST at $\epsilon = 0.1$ the separation is sharpest: S2 99.4, confidence 32.6, Mahalanobis 71.1. S3 and S4 track S2 within a point wherever S2 is strong; S4 also keeps most of confidence's clean advantage and is the safer default.

Detection: parity with the strongest detector. For separating attacked from clean inputs (Table 4), S2 reaches 83 on CIFAR-10 (PGD, C&W, AutoAttack), 86–89 on GTSRB and 99.5 on MNIST; Mahalanobis is higher on CIFAR-10 (87–98) and comparable on GTSRB and MNIST. But Mahalanobis cannot do selective prediction: it detects that an input was attacked, not whether the teacher survived, ranking GTSRB teacher errors at 47 (PGD) and 57 (C&W) against 96 for S2 and wasting 7–16% of human queries on inputs the teacher had right where S2 wastes none (Table 2). The two signals are complementary.

Student-only measures are blind to transfer-resistant attacks. S5–S7 sit at detection AUROC 43–58 under every non-adaptive attack on every dataset. The header rows of Table 1 explain why: attacks crafted for the teacher barely reach the students (CIFAR-10 students hold 75–76% under PGD, C&W and AutoAttack against 76.0% clean; GTSRB 84–86% against 90.7%; MNIST 98.8% against 99.1%), so their internal uncertainty does not move. On GTSRB, where part of the attack transfers, S5 and S7 rank teacher errors reasonably (87–90) while still failing to detect the attack. The same transfer failure makes the ensemble a usable **fallback**: where the teacher is at 0–9%, the student majority is at 75% (CIFAR-10), 85% (GTSRB) and 99% (MNIST) (Table 3).

Adaptive attack. When the adversary attacks teacher and students jointly while suppressing disagreement, S2's detection AUROC falls to 75 (CIFAR-10), 73 (GTSRB) and 97 (MNIST), and fallback accuracy to 56%, 61% and 87%. This is degradation, not collapse: confidence-based scores stay at 0–16, and on GTSRB S2 still ranks teacher errors at 84 (Table 1, Adapt). Two side effects: entropy (S5) becomes the best GTSRB error ranker (88.0) and detects the attack on MNIST at 88.7, because an attacker who must perturb the students to evade disagreement thereby makes them uncertain; and Mahalanobis, outside the attacker's objective, is unaffected (82–100). Nor can the attacker tune its way out: raising λ lowers disagreement's detection AUROC but leaves its error ranking at 82–84 and restores teacher accuracy, because budget spent on agreement is not spent on fooling the teacher (Appendix H).

5 Discussion and Limitations

What the results establish. S2 and its combinations S3/S4 are the only rules that rank teacher errors well on every dataset where the students are adequate and under every non-adaptive attack, and their advantage is largest exactly where confidence is worst; the mechanism is imperfect transferability between the teacher and heterogeneous robust students (Tramèr et al., 2017), which the header rows of Table 1 measure directly. Every confidence-derived score, including MC dropout and a selector learned on clean data, becomes an anti-signal under multi-step attacks, sharpening the failure reported by Stutz et al. (2020): attacks make wrong predictions look *more* trustworthy than clean ones. Student-only measures (entropy, mutual information, vote margin) move only when the attack reaches the students, so for teacher-targeted attacks they are best used through the teacher–student comparison rather than as stand-alone detectors, which is what our disagreement rule does. Feature squeezing is a strong competitor at small budgets ($\epsilon = 2/255$, Appendix G), where a median filter undoes most of the perturbation, and loses that advantage at $8/255$; it and Mahalanobis fail in different regimes from disagreement and are natural to combine.

Threat model. Our main attacks are white-box on the teacher and reach the students only by transfer, the setting in which disagreement is strongest. One adaptive attack with a λ sweep degrades detection without collapsing error ranking, but a full adaptive evaluation (Tramer et al., 2020) with other objectives (a common wrong class across all models, attacks on the students' features) and more steps is needed before the adaptive-robustness claim is quantitative. We consider ℓ_∞ adversaries only, and sweep the budget on CIFAR-10, MNIST and Tiny ImageNet but not GTSRB.

Clean cost and student capacity. Disagreement's clean false-alarm rate grows with the student–teacher accuracy gap (0.5%, 7%, 23%, 67% at gaps of 0.5, 4, 19, 44 points on MNIST, GTSRB, CIFAR-10 and Tiny ImageNet), which fixes the method's operating condition: the students should be within roughly 20 points of the teacher. Our lightweight Tiny ImageNet students fall outside it, which bounds the regime rather than contradicting the results within it; larger students (which we did not tune), or combining disagreement with confidence (S4) or a feature-space detector, would lower the cost.

Oracle and scope. Oracle accuracy assumes a perfect human and upper-bounds what an imperfect, overloaded reviewer would attain. Results are single-seed, and extending the framework to detection, segmentation or sequence tasks requires diverse student architectures for structured outputs.

Which ingredient matters. Appendix D (Table 5) removes one ingredient at a time, comparing each ensemble against the same teacher on identical inputs. Two ingredients are load-bearing. *Distillation* keeps the students accurate enough to be a reference: without it (D) mean student accuracy falls from 74.5 to 51.6 on CIFAR-10 and from 99.0 to 55.3 on MNIST, and on MNIST this destroys the signal outright (clean false-deferral rate 78.8%, disagreement at chance), the same capacity limit reached from the other side in Section 5. *Adversarial training* makes the ensemble a *robust* reference: without it (C) attacks transfer to the students far more (PGD transfer $1.1\% \to 12.1\%$ on CIFAR-10, $6.4\% \to 29.3\%$ on GTSRB) and under the adaptive attack C collapses, its fallback accuracy falling from 55.5% to 0.7% (CIFAR-10) and 87.0% to 0.2% (MNIST) while its disagreement detection drops to chance. *Architectural* diversity, by contrast, is a minor factor: a same-architecture, six-seed ensemble (B) matches the diverse one within about two points on every deferral metric, since independent seeds already limit transfer. The method's robustness thus comes chiefly from adversarial training and distillation, with architectural diversity a modest addition.

6 Conclusion

We paired a standard-trained teacher with a small ensemble of distilled, adversarially trained, architecturally diverse student models and used their disagreement to decide when to defer to a human. Across CIFAR-10, GTSRB and MNIST under FGSM, PGD, C&W and AutoAttack, disagreement is the strongest signal for deciding which teacher predictions to trust, matching the best adversarial detector on attack detection while that detector cannot rank teacher errors; confidence-based scores become anti-signals; and the ensemble doubles as a fallback with 75–99% accuracy where the teacher is at 0–9%.

Ethics Statement

This paper studies robust decision-making in human-in-the-loop machine learning systems under adversarial conditions. The primary goal of the proposed framework is to improve the reliability and safety of deployed AI systems by enabling principled deferral to human experts when model predictions may be compromised by adversarial manipulation or uncertainty. By reducing over-reliance on automated predictions in high-risk scenarios, this work has the potential to positively impact safety-critical applications such as autonomous systems, medical decision support, and security-sensitive classification tasks. Our work does not introduce new attack methods beyond a straightforward ensemble-aware adaptive objective used to test our own defense, and focuses on defensive strategies that emphasize human oversight rather than increased automation. The proposed framework assumes the availability of qualified human experts when deferral occurs, and we caution that the oracle accuracies we report are upper bounds that would not be attained with imperfect or overloaded human reviewers. All experiments use publicly available benchmark datasets (MNIST, CIFAR-10, GTSRB, Tiny ImageNet) that contain no personally identifying information and involve no human subjects.

Reproducibility Statement

We have taken the following steps to make our results reproducible. The teacher architectures and training objective are specified in Section 3.1; the student architectures, their structural differences, initialization, and the complete distillation and adversarial-training recipe (temperature, distillation weight, clean/adversarial loss mixing, distillation target, PGD parameters and ϵ warm-up, optimizer, learning-rate schedule, batch size, epoch budget, and checkpoint-selection criterion on a held-out training slice) are given in Section 3.2. All attack budgets are stated in $[0, 1]$ pixel space with per-model normalization folded into the model, and every checkpoint carries a record of its training-time preprocessing and final clean and robust accuracy, which the evaluation harness verifies before use. All four datasets are public benchmarks and the preprocessing we apply is stated in Section 3.5. Every deferral rule, baseline, and metric is defined in Sections 3.4–3.5, and the attack configurations in Section 4. Appendix E reports the fixed-threshold autonomous-accuracy and query-rate breakdown. Anonymized source code for training the teacher and students, generating the adversarial test sets, and reproducing all tables is provided as supplementary material.

AI Use Statement

We used generative AI for writing assistance and editing. We did not use it for research ideation or method design. All AI-assisted text were reviewed and verified by the authors, who take full responsibility for this work.

References

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pp. 274–283. PMLR, 2018.

Mohammad Azizmalayeri, Arshia Soltani Moakhar, Arman Zarei, Reihaneh Zohrabi, Mohammad Manzuri, and Mohammad Hossein Rohban. Your out-of-distribution detection method is not robust! *Advances in Neural Information Processing Systems*, 35:4887–4901, 2022.

Steven Basart, Mazeika Mantas, Mostajabi Mohammadreza, Steinhardt Jacob, and Song Dawn. Scaling out-of-distribution detection for real-world settings. In *International Conference on Machine Learning*, 2022.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. Ieee, 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*, pp. 2206–2216. PMLR, 2020.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International conference on machine learning*, pp. 2151–2159. PMLR, 2019.

Micah Goldblum, Liam Fowl, Soheil Feizi, and Tom Goldstein. Adversarially robust distillation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 3996–4003, 2020.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv preprint arXiv:1810.12715*, 2018.

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Erik Jones, Shiori Sagawa, Pang Wei Koh, Ananya Kumar, and Percy Liang. Selective classification can magnify disparities across groups. *arXiv preprint arXiv:2010.14134*, 2020.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Yann Le, Xuan Yang, et al. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11966–11976. IEEE, 2022.

David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: improving fairness and accuracy by learning to defer. *Advances in neural information processing systems*, 31, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Yannis Montreuil, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. Adversarial robustness in two-stage learning-to-defer: Algorithms and guarantees. *arXiv preprint arXiv:2502.01027*, 2025.

Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. Human-in-the-loop machine learning: a state of the art: E. mosqueira-rey et al. *Artificial Intelligence Review*, 56(4):3005–3054, 2023.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *International conference on machine learning*, pp. 7076–7087. PMLR, 2020.

Nastaran Okati, Abir De, and Manuel Rodriguez. Differentiable learning under triage. *Advances in Neural Information Processing Systems*, 34:9140–9151, 2021.

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.

Tianyu Pang, Kun Xu, Chao Du, Ning Chen, and Jun Zhu. Improving adversarial robustness via promoting ensemble diversity. In *International Conference on Machine Learning*, pp. 4970–4979. PMLR, 2019.

Lewis Smith and Yarin Gal. Understanding measures of uncertainty for adversarial example detection. *arXiv preprint arXiv:1803.08533*, 2018.

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.

David Stutz, Matthias Hein, and Bernt Schiele. Confidence-calibrated adversarial training: Generalizing to unseen attacks. In *International conference on machine learning*, pp. 9155–9166. PMLR, 2020.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *Advances in neural information processing systems*, 33:1633–1645, 2020.

Rajeev Verma and Eric Nalisnick. Calibrated learning to defer with one-vs-all classifiers. In *International Conference on Machine Learning*, pp. 22184–22202. PMLR, 2022.

Bryan Wilder, Eric Horvitz, and Ece Kamar. Learning to complement humans. *arXiv preprint arXiv:2005.00582*, 2020.

Table 4: Adversarial-detection AUROC (↑): separation of attacked from clean inputs by each score. Values near 0 mean the score separates the two almost perfectly *in the wrong direction* (attacked inputs look more trustworthy than clean ones); values near 50 mean the score does not react to the attack at all. Best per column in **bold**.

	CIFAR-10					GTSRB					MNIST					Tiny ImageNet				
	FGSM	PGD	C&W	AA	Adapt	FGSM	PGD	C&W	AA	Adapt	FGSM	PGD	C&W	AA	Adapt	FGSM	PGD	C&W	AA	Adapt
S1 Low Confidence	85.2	0.2	0.0	12.6	0.7	88.7	37.0	38.5	35.0	49.0	98.0	11.9	4.4	18.5	16.3	59.0	16.9	10.0	40.0	17.0
S2 Disagreement	68.2	83.1	83.0	83.8	75.4	78.5	86.7	85.7	**88.9**	73.2	94.7	99.5	99.6	99.4	97.3	56.8	60.3	59.6	60.4	57.4
S3 S1 and S2	71.8	71.4	64.5	74.8	66.4	79.6	83.6	82.7	85.9	71.9	94.7	99.2	99.2	99.1	97.0	59.0	35.5	31.2	50.6	34.1
S4 S1 or S2	79.7	79.3	79.1	82.3	65.8	89.8	87.7	86.3	86.5	70.7	99.4	99.5	99.5	99.5	95.5	57.7	59.9	59.0	60.5	56.6
S5 Predictive Entropy	50.4	50.7	50.8	50.7	59.3	57.3	56.6	56.8	56.2	67.1	46.5	42.9	43.1	44.1	88.7	50.5	50.6	50.7	50.6	52.4
S6 Vote Margin	50.1	50.1	50.3	50.2	54.6	53.1	51.9	52.3	51.9	56.4	50.4	50.2	50.2	50.3	56.5	50.0	50.0	50.1	50.2	51.1
S7 Mutual Information	50.3	50.5	50.6	50.5	58.4	56.3	55.3	55.4	55.0	66.1	66.5	58.6	58.6	62.0	81.9	49.5	49.3	49.2	49.4	47.7
Max-Logit	84.9	0.1	0.0	10.3	0.6	87.7	32.1	32.5	32.0	43.8	95.1	5.4	1.1	15.0	9.3	62.6	14.2	8.7	37.8	14.2
Teacher Entropy	85.3	0.2	0.0	12.4	0.7	88.6	36.4	37.7	34.5	48.3	97.7	10.4	3.3	17.7	14.8	61.4	16.5	9.8	39.5	16.5
Energy	84.9	0.1	0.0	10.3	0.6	87.7	32.1	32.5	32.0	43.7	94.4	4.7	0.9	14.7	8.6	65.2	14.6	9.1	37.3	14.7
ODIN	87.9	0.2	0.0	7.0	0.8	88.8	33.7	34.1	33.1	45.2	95.9	3.6	0.7	14.5	7.5	**66.7**	7.9	5.1	31.7	8.8
Mahalanobis	**89.9**	**92.5**	**98.0**	**86.9**	**87.5**	**90.8**	**87.5**	**89.4**	82.0	**82.3**	**100.0**	**99.8**	**99.7**	**99.9**	**99.8**	52.7	64.3	78.1	63.6	62.0
MC Dropout	85.1	0.2	0.0	12.4	0.8	87.0	38.6	39.6	36.2	49.4	97.6	10.6	3.6	17.8	15.1	61.4	16.8	10.1	39.6	16.8
Feature Squeezing	72.0	84.6	85.3	84.2	78.9	87.2	60.6	60.9	62.1	67.3	97.9	98.4	98.9	98.1	98.4	60.5	**80.5**	**80.7**	**77.5**	**78.6**
Learned Deferral	84.5	0.1	0.0	10.8	0.7	89.7	53.5	51.3	55.2	60.3	96.7	44.3	52.1	67.1	71.4	56.8	45.6	44.2	56.4	44.5

Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International conference on machine learning*, pp. 5286–5295. PMLR, 2018.

Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017.

A Additional Tables

Table 4 reports adversarial-detection AUROC for every score and condition; Table 2 reports oracle accuracy at 25% and 50% human-query budgets with wasted slots; Table 3 reports the teacher and student-majority accuracies under every condition.

B Student Architectures

All six students use ReLU activations, end with adaptive average pooling to a 4×4 grid, and feed a two-layer fully-connected head ($C \cdot 16 \to 256 \to K$). The convolutional trunks differ as follows.

1. **Depth and block structure**: S_1 and S_2 use 5 plain 3×3 convolutions; S_3 a shallower 4-layer trunk; S_4 7 convolutions in two bottleneck blocks ($3\times3 \to 1\times1$ compress $\to 3\times3$); S_5 a 6-layer trunk in which every second convolution is dilated (dilation 2); S_6 the shallowest trunk, 3 convolutions.
2. **Downsampling**: MaxPool (S_1, S_3, S_4), strided 3×3 convolutions (S_2, S_6), or AvgPool (S_5).
3. **Kernel sizes**: 3×3 throughout (S_1, S_2, S_3); mixed 3×3 and 1×1 (S_4); 3×3 with alternating dilation (S_5); a 5×5 stem followed by 3×3 (S_6).
4. **Channel widths**: $3 \to 32 \to 64 \to 128$ (S_1, S_2); $3 \to 64 \to 128$ (S_3); $3 \to 32 \to 16 \to 32 \to 64 \to 32 \to 64 \to 128$ (S_4); $3 \to 48 \to 96 \to 192$ (S_5); $3 \to 64 \to 128 \to 128$ (S_6).

C Datasets

MNIST: 70,000 grayscale handwritten digits (60,000 train, 10,000 test), 10 classes; the grayscale channel is replicated to three and images resized to 32×32. **CIFAR-10**: 60,000 32×32 color images in 10 classes (50,000 / 10,000). **GTSRB**: over 50,000 color images of 43 traffic-sign classes in real driving conditions (39,209 / 12,630), resized to 32×32; its larger, imbalanced label space is characteristic of safety-critical applications. **Tiny ImageNet**: a 200-class ImageNet subset at 64×64 (100,000 train, 10,000 validation).

Table 5: **Ensemble ablation: which ingredient matters.** Each ensemble is compared against the same teacher on identical attacked inputs. A = diverse architectures + adversarial training + distillation (the paper's ensemble); B = one architecture, six seeds, adversarial training + distillation (isolates *architectural* diversity from ensembling); C = diverse, *standard* training (no adversarial branch); D = diverse, adversarial training, *no distillation*. Stud. = mean clean accuracy of the individual student models; FDR = clean false-deferral rate of disagreement (↓); Err / Or@50 = error-detection AUROC and oracle accuracy at a 50% budget for disagreement (S2); Tr = attack transfer rate to the ensemble (↓); Fb = student-majority fallback accuracy; Adv = adversarial-detection AUROC of disagreement. FGSM carries the error ranking (teacher partially broken); PGD and the ensemble-aware adaptive attack carry transfer, fallback and detection. All values %.

		Clean		FGSM		PGD			Adaptive		
Data	Ens.	Stud.	FDR	Err	Or@50	Tr	Fb	Adv	Tr	Fb	Adv
CIFAR-10	A	74.5	21.9	86.9	83.3	1.1	75.0	83.0	26.8	55.5	75.3
	B	74.0	22.8	86.4	83.2	1.0	74.2	82.4	27.6	54.2	74.5
	C	85.9	8.9	83.0	84.2	12.1	77.8	87.1	99.2	0.7	46.4
	D	51.6	26.0	84.5	82.2	1.0	71.1	80.7	17.4	59.3	76.6
GTSRB	A	89.9	7.1	96.6	77.1	6.4	84.9	86.6	33.3	60.4	73.2
	B	88.8	8.3	96.3	77.0	6.9	83.4	85.9	34.8	58.3	72.3
	C	95.4	1.5	87.7	77.1	29.3	68.1	78.3	64.0	34.6	59.0
	D	85.3	11.3	96.5	76.9	5.3	82.3	85.1	32.3	58.7	68.7
MNIST	A	99.0	0.6	99.0	59.9	0.3	98.9	99.6	12.3	87.0	97.3
	B	69.9	0.6	98.9	59.9	0.3	98.9	99.5	10.2	88.9	98.3
	C	99.5	0.2	94.8	59.8	1.3	98.3	99.6	99.8	0.2	55.2
	D	55.3	78.8	50.0	54.8	0.1	21.1	60.6	1.6	20.8	60.6

D Ensemble Ablation

Table 5 isolates the contribution of each ingredient of the student ensemble. See Section 5 for discussion.

E Fixed-Threshold Framework Breakdown

For comparability with prior selective-prediction work, Table 6 reports the fixed-threshold form of each strategy ($\tau = \tau_H = \tau_V = 0.5$, $\tau_I = 0.3$): the autonomous framework accuracy on the subset where the framework predicts (Frmwk), the human query rate (Query), and the resulting oracle accuracy with a perfect human (Oracle). Reading Frmwk and Query together is essential: an Oracle value near 100% achieved by deferring nearly 100% of inputs is qualitatively different from the same Oracle achieved at high autonomous coverage, which is why the main text reports threshold-free metrics. Two things the table shows that the main text relies on. First, the old headline: S5 reaches 100% oracle accuracy on every Tiny ImageNet condition by deferring every input, and 99.2% on clean CIFAR-10 by deferring half of them; neither is a selective-prediction result. Second, the binary rule S2 has a single operating point, which under attack sits at a high query rate (85–100% on PGD) because most inputs genuinely need deferral there; Table 2 compares rules at equal query rates instead.

F Selective Risk Under Attack

Table 7 reports AURC (↓, %) on the conditions where the teacher retains non-zero accuracy, complementing the AUROC values of Table 1. Risk–coverage curves for every condition are provided in the supplementary material.

Table 6: Fixed-threshold metrics (%) at $\tau = \tau_H = \tau_V = 0.5$, $\tau_I = 0.3$, full attack budgets. Query = human query rate; Frmwk = accuracy on the autonomously answered subset ("—" when nothing is answered); Oracle = overall accuracy with a perfect human. Note S5 on Tiny ImageNet: 100% oracle accuracy by deferring 100% of inputs, the artefact that motivates the threshold-free metrics in the main text. Binary rules (S2) show as a single operating point; AutoAttack and adaptive conditions are in the released results.

		Clean			FGSM			PGD			C&W		
Dataset	Strategy	Query	Frmwk	Oracle	Query	Frmwk	Oracle	Query	Frmwk	Oracle	Query	Frmwk	Oracle
CIFAR-10	S1	0.6	95.1	95.1	3.8	38.1	40.5	0.0	0.0	0.0	0.0	0.0	0.0
	S2	23.8	97.3	97.9	60.1	80.9	92.4	90.0	0.0	90.0	89.8	0.0	89.8
	S3	0.4	95.0	95.0	3.0	38.4	40.2	0.0	0.0	0.0	0.0	0.0	0.0
	S4	23.9	97.4	98.0	60.9	81.1	92.6	90.0	0.0	90.0	89.8	0.0	89.8
	S5	51.3	98.3	99.2	51.5	47.9	74.7	52.0	0.0	52.0	52.1	0.0	52.1
	S6	13.0	95.5	96.1	13.1	39.2	47.2	13.1	0.0	13.1	13.1	0.0	13.1
	S7	0.1	94.8	94.8	0.1	37.3	37.4	0.2	0.0	0.2	0.2	0.0	0.2
GTSRB	S1	3.4	97.0	97.1	24.9	31.4	48.4	3.5	7.9	11.1	1.8	8.8	10.5
	S2	11.3	99.2	99.3	68.3	85.8	95.5	84.7	57.0	93.4	82.8	53.3	92.0
	S3	2.7	96.8	96.9	20.7	34.3	47.9	2.4	9.0	11.1	1.4	9.3	10.5
	S4	12.0	99.4	99.5	72.4	85.5	96.0	85.8	53.7	93.4	83.3	51.9	92.0
	S5	26.5	99.2	99.5	36.2	42.0	63.0	34.6	13.4	43.4	35.2	14.2	44.4
	S6	5.8	95.9	96.1	9.2	30.0	36.4	8.2	9.5	16.9	8.7	10.0	17.8
	S7	0.8	94.9	95.0	0.9	27.5	28.2	0.9	8.8	9.6	0.8	9.2	10.0
MNIST	S1	0.2	99.7	99.7	3.1	10.0	12.8	0.0	0.0	0.0	0.0	0.0	0.0
	S2	0.7	99.7	99.7	90.1	97.6	99.8	99.8	0.0	99.8	99.9	0.0	99.9
	S3	0.1	99.7	99.7	2.9	10.1	12.8	0.0	0.0	0.0	0.0	0.0	0.0
	S4	0.8	99.8	99.8	90.3	97.7	99.8	99.8	0.0	99.8	99.9	0.0	99.9
	S5	15.0	100.0	100.0	18.6	9.2	26.1	17.8	0.0	17.8	17.7	0.0	17.7
	S6	0.6	99.6	99.6	1.2	9.9	11.0	0.9	0.0	0.9	0.9	0.0	0.9
	S7	0.0	99.6	99.6	0.0	9.9	9.9	0.0	0.0	0.0	0.0	0.0	0.0
Tiny ImageNet	S1	27.1	85.1	89.1	34.5	9.9	41.0	2.3	0.0	2.3	0.1	0.0	0.1
	S2	72.5	87.5	96.5	86.2	42.9	92.1	93.1	0.0	93.1	91.8	0.0	91.8
	S3	24.5	83.6	87.6	31.4	11.9	39.6	2.3	0.0	2.3	0.1	0.0	0.1
	S4	75.1	92.3	98.1	89.3	39.5	93.5	93.2	0.0	93.2	91.8	0.0	91.8
	S5	100.0	—	100.0	100.0	—	100.0	100.0	—	100.0	100.0	—	100.0
	S6	61.4	76.5	90.9	61.6	14.7	67.2	61.4	0.0	61.4	61.6	0.0	61.6
	S7	7.5	69.7	72.0	6.9	8.0	14.3	6.9	0.0	6.9	6.9	0.0	6.9

Table 7: AURC (↓, %). Lower is better; best per column in **bold**.

	GTSRB						CIFAR-10		MNIST		Tiny	
	Clean	FGSM	PGD	C&W	AA	Adapt	Clean	FGSM	Clean	FGSM	Clean	FGSM
S1 Low Confidence	0.31	67.5	98.7	98.5	95.8	95.2	**0.64**	59.4	0.11	82.5	9.55	93.4
S2 Disagreement	0.99	41.7	**74.7**	**74.2**	79.5	72.0	3.12	36.1	0.40	67.9	20.9	**79.6**
S3 S1 and S2	0.77	**41.5**	74.9	74.8	**79.3**	71.3	3.13	**35.9**	0.36	67.9	13.9	80.7
S4 S1 or S2	0.42	43.8	80.4	80.7	81.9	83.3	1.17	37.8	0.11	**67.6**	18.1	82.3
S5 Predictive Entropy	0.64	45.8	79.3	78.0	82.9	**65.3**	2.03	50.0	**0.01**	88.5	22.1	82.6
S6 Vote Margin	2.65	66.7	89.9	89.6	92.5	81.0	3.93	59.7	0.23	90.3	25.5	87.6
S7 Mutual Information	0.95	46.1	79.8	78.5	83.4	67.0	2.36	51.6	0.03	96.2	33.6	94.2
Mahalanobis	0.32	60.0	93.3	90.8	79.0	86.4	0.69	53.0	0.02	84.8	10.9	85.2
Feature Squeezing	0.43	62.2	96.8	96.4	87.1	90.8	1.33	44.7	0.04	80.5	14.2	83.6
Learned Deferral	**0.28**	49.7	97.0	96.2	90.5	90.4	0.77	61.0	**0.01**	69.1	**9.26**	88.2

G BUDGET SWEEPS

At the standard budgets ($8/255$; 0.3 on MNIST) every multi-step attack drives the CIFAR-10, MNIST and Tiny ImageNet teachers to 0%, so error ranking is defined only under FGSM. This appendix repeats the full evaluation at reduced budgets with the same teachers, students, baselines and attacks: CIFAR-10 at $2/255$ and $4/255$ (Tables 8, 9), MNIST at 0.1 and 0.2 (Tables 10, 11), and Tiny ImageNet at $2/255$ and $4/255$ (Tables 12, 13). The partially-broken columns from these runs are the ones shown in Table 1. Budgets follow the standard convention for each dataset; for completeness we also evaluated MNIST at $8/255$, where the teacher retains 98.4–98.7% accuracy under every attack

and no score separates attacked from clean inputs above AUROC 80 (most near 50): $8/255$ is not a meaningful budget on a near-binary image dataset, which is why MNIST is evaluated at 0.3.

CIFAR-10. At $2/255$ the teacher retains 58.6% (FGSM), 7.9% (PGD), 5.7% (C&W), 3.4% (AutoAttack) and 12.8% (adaptive), so selective prediction is defined under every attack, as on GTSRB in the main text. At $4/255$ the teacher survives PGD, C&W and AutoAttack on only 14, 3 and 1 of $8{,}000$ images; error-ranking AUROC on those columns rests on a handful of positives and is reported in grey for completeness only.

Three observations. First, the GTSRB pattern reproduces on CIFAR-10: at $2/255$ disagreement ranks teacher errors at AUROC 93.4 (PGD), 93.2 (C&W) and 94.7 (AutoAttack) where confidence scores 6.1, 4.8 and 39.8 and Mahalanobis 28.2, 32.5 and 91.9; at a 50% query budget under PGD, disagreement recovers 57.8% oracle accuracy against 50.0% for confidence. Second, the confidence inversion is already complete at $2/255$: all confidence-family scores rank wrong predictions as more trustworthy than right ones under every multi-step attack. Third, feature squeezing is a strong competitor at small budgets, tied with disagreement under FGSM (85.1 vs. 85.0) and PGD (89.6 vs. 93.4) and ahead under AutoAttack (99.0 vs. 94.7), because the median filter removes most of a $2/255$ perturbation; by $8/255$ its error-ranking AUROC under FGSM has fallen to 76.7 (Table 1). Student-only measures remain at detection AUROC ≈ 50 at every budget, and the ensemble's accuracy is 75–76% under every non-adaptive attack at every budget.

Table 8: CIFAR-10 at $\epsilon = 2/255$. Err = error-detection AUROC; Adv = adversarial-detection AUROC; Or@50 = oracle accuracy at a 50% query budget. Teacher accuracy: FGSM 58.6, PGD 7.9, C&W 5.7, AA 3.4, Adapt 12.8. Student majority: 75.8 / 75.7 / 75.7 / 75.8 / 72.7.

	FGSM			PGD			C&W			AutoAttack			Adaptive		
	Err	Adv	Or@50	Err	Adv	Or@50	Err	Adv	Or@50	Err	Adv	Or@50	Err	Adv	Or@50
S1 Low Confidence	56.6	79.1	83.1	6.1	27.2	50.0	4.8	22.0	50.0	39.8	32.0	50.6	9.0	31.7	50.0
S2 Disagreement	85.0	58.6	93.8	**93.4**	79.1	57.8	**93.2**	79.9	55.7	94.7	81.5	53.4	**91.0**	75.6	62.4
S3 S1 and S2	**85.2**	60.3	93.8	92.5	73.0	57.8	92.4	73.0	55.6	94.7	76.5	53.4	89.6	70.4	62.3
S4 S1 or S2	84.0	73.8	**94.0**	88.0	78.3	57.8	87.7	78.1	55.6	93.2	79.6	53.4	84.2	75.7	62.5
S5 Predictive Entropy	74.6	50.1	88.1	82.6	50.2	57.2	84.6	50.3	55.2	88.7	50.2	53.3	79.6	52.2	61.1
S6 Vote Margin	60.9	50.0	83.0	62.0	50.1	55.3	62.1	50.1	53.7	62.9	50.1	52.1	62.4	51.0	58.2
S7 Mutual Information	69.9	50.1	86.8	77.3	50.2	56.6	80.2	50.2	55.0	84.8	50.2	53.1	73.2	52.0	60.0
Max-Logit	52.2	79.7	81.2	5.2	25.9	50.0	4.8	20.1	50.0	35.0	29.6	50.3	7.6	30.4	50.0
Teacher Entropy	56.4	79.2	83.1	6.0	27.0	50.0	4.8	21.8	50.0	39.5	31.8	50.6	8.9	31.6	50.0
Energy	52.0	79.7	81.2	5.2	25.9	50.0	4.8	20.1	50.0	35.0	29.6	50.3	7.6	30.4	50.0
ODIN	37.2	81.2	75.1	4.7	27.9	50.0	4.6	21.6	50.0	25.6	27.2	50.2	6.6	32.8	50.0
Mahalanobis	65.7	77.1	85.1	28.2	76.3	51.7	32.5	77.7	51.5	91.9	75.0	53.3	40.9	74.0	54.8
MC Dropout	56.2	79.0	82.9	5.9	27.1	50.0	4.8	22.0	50.0	37.9	31.9	50.4	8.8	31.7	50.0
Feature Squeezing	85.1	66.3	92.8	89.6	**87.4**	**57.9**	89.6	**88.3**	**55.7**	**99.0**	**87.6**	**53.4**	89.4	**85.0**	**62.6**
Learned Deferral	54.4	79.3	81.8	5.5	26.1	50.0	4.9	20.2	50.0	37.5	30.2	50.5	8.1	30.9	50.0

MNIST. At $\epsilon = 0.1$ the teacher survives PGD at 25.5% and C&W at 22.1%, and disagreement ranks its errors at AUROC 99.4 where confidence scores 30–33 and Mahalanobis 67–71; at a 50% budget under PGD disagreement recovers 75.5% against 59.1% for confidence. At 0.2 the teacher is already at 0% under every multi-step attack. The students hold 98–99% throughout.

Tiny ImageNet. At $2/255$ the teacher survives PGD at 8.2%. With students at 26%, disagreement ranks teacher errors at 76–81, above confidence (18–76) but below feature squeezing (88–96) and, under AutoAttack, Mahalanobis (97). This is the capacity limit of Section 5 at a budget where the ranking question is defined: the students are too weak for disagreement to be the best signal here.

H Adaptive Attack: λ Sweep

The ensemble-aware adaptive attack of Section 4 maximises $\mathrm{CE}(f_\theta(x'), y) + \sum_j \mathrm{CE}(f_{\phi_j}(x'), y) - \lambda \sum_j \mathrm{KL}\big(f_\theta(x') \,\|\, f_{\phi_j}(x')\big)$: the first two terms fool teacher and students, the third pulls the students toward the teacher's output so that they agree on the wrong answer. λ sets how much of the $\epsilon = 8/255$ budget is spent on suppressing disagreement rather than on misclassification. Table 14 reports GTSRB at $\lambda \in \{0.3, 1, 3, 10\}$ with 20 PGD steps; $\lambda = 1$ is the setting used in the main text.

Table 9: CIFAR-10 at $\epsilon = 4/255$. Columns as in Table 8. Teacher accuracy: FGSM 49.0, PGD 0.18, C&W 0.04, AA 0.01, Adapt 1.4. Student majority: 75.6 / 75.7 / 75.6 / 75.5 / 68.1. Err values in grey rest on ≤ 14 surviving teacher predictions and are not interpretable.

	FGSM			PGD			C&W			AutoAttack			Adaptive		
	Err	Adv	Or@50	Err	Adv	Or@50	Err	Adv	Or@50	Err	Adv	Or@50	Err	Adv	Or@50
S1 Low Confidence	56.5	81.6	77.8	0.4	3.5	50.0	0.0	1.3	50.0	29.5	16.0	50.0	1.5	7.3	50.0
S2 Disagreement	86.3	62.6	92.3	94.8	82.8	50.2	94.6	82.7	50.0	95.2	83.3	50.0	**91.3**	79.3	51.3
S3 S1 and S2	**86.7**	65.0	92.3	94.7	72.8	50.2	93.3	71.4	50.0	95.2	75.1	50.0	90.9	70.6	51.3
S4 S1 or S2	85.5	76.4	**92.3**	89.6	79.0	50.2	89.2	78.6	50.0	93.8	81.5	50.0	83.2	73.8	51.3
S5 Predictive Entropy	71.2	50.2	81.9	88.2	50.5	50.2	95.6	50.5	50.0	96.9	50.4	50.0	83.7	54.8	51.3
S6 Vote Margin	58.9	50.1	77.6	63.6	50.1	50.2	63.6	50.2	50.0	63.6	50.1	50.0	62.7	52.1	51.0
S7 Mutual Information	67.0	50.1	81.0	80.9	50.4	50.2	92.1	50.4	50.0	95.7	50.3	50.0	76.9	54.4	51.1
Max-Logit	52.4	81.9	76.3	0.3	3.2	50.0	0.0	1.1	50.0	22.8	14.2	50.0	1.2	7.0	50.0
Teacher Entropy	56.3	81.7	77.7	0.4	3.5	50.0	0.0	1.2	50.0	29.0	15.8	50.0	1.5	7.3	50.0
Energy	52.1	81.9	76.2	0.3	3.2	50.0	0.0	1.1	50.0	22.8	14.2	50.0	1.2	7.0	50.0
ODIN	41.5	84.3	71.6	0.3	3.9	50.0	0.0	1.5	50.0	12.7	10.4	50.0	1.1	8.0	50.0
Mahalanobis	63.0	84.2	79.3	29.9	84.3	50.0	18.9	**91.4**	50.0	66.5	81.3	50.0	39.7	80.3	50.5
MC Dropout	56.0	81.5	77.6	0.4	3.6	50.0	0.0	1.3	50.0	26.9	15.8	50.0	1.5	7.4	50.0
Feature Squeezing	81.2	69.0	86.9	86.3	**88.3**	50.2	81.0	88.4	50.0	99.9	**88.5**	50.0	85.5	**86.1**	**51.4**
Learned Deferral	54.5	81.6	76.9	0.3	3.3	50.0	0.0	1.1	50.0	25.4	14.6	50.0	1.3	7.0	50.0

Table 10: MNIST at $\epsilon = 0.1$. Columns as in Table 8. Teacher accuracy: FGSM 94.8, PGD 25.5, C&W 22.1, AA 2.4, Adapt 89.3; student majority 98.8 / 98.8 / 98.8 / 98.9 / 97.8. At this budget the adaptive attack, which must split 0.1 across seven robust models, barely affects the teacher, so its column is uninformative here.

	FGSM			PGD			C&W			AutoAttack			Adaptive		
	Err	Adv	Or@50	Err	Adv	Or@50	Err	Adv	Or@50	Err	Adv	Or@50	Err	Adv	Or@50
S1 Low Confidence	86.6	94.5	99.5	32.6	75.0	59.1	29.7	68.2	56.4	74.5	71.7	**52.4**	52.2	67.3	95.0
S2 Disagreement	92.6	52.0	99.6	**99.4**	86.5	**75.5**	**99.4**	88.1	**72.1**	99.6	98.1	**52.4**	92.9	54.4	99.2
S3 S1 and S2	92.7	52.0	99.6	**99.4**	86.4	**75.5**	**99.4**	88.0	**72.1**	99.6	98.0	**52.4**	92.9	54.4	99.3
S4 S1 or S2	95.4	94.3	99.8	98.8	**98.3**	**75.5**	98.8	97.9	**72.1**	99.6	98.0	**52.4**	89.5	70.6	98.8
S5 Predictive Entropy	90.0	38.2	99.6	64.5	41.5	67.0	61.5	42.2	63.8	65.1	41.5	51.7	90.5	71.1	99.5
S6 Vote Margin	60.5	50.1	97.8	51.1	50.2	62.7	51.1	50.2	61.4	50.9	50.1	51.1	58.6	50.7	95.3
S7 Mutual Information	83.9	55.8	99.4	57.7	56.3	64.0	51.6	56.4	60.9	28.1	56.4	50.2	86.7	55.7	99.4
Max-Logit	86.0	94.0	99.5	29.1	70.8	57.6	27.5	62.9	55.0	70.3	65.4	**52.4**	48.4	67.0	94.6
Teacher Entropy	86.5	94.4	99.5	31.8	74.2	58.8	29.3	67.1	56.0	73.3	70.6	**52.4**	51.4	67.2	94.9
Energy	85.8	93.9	99.5	28.5	70.2	57.4	27.3	62.3	54.8	69.8	64.6	**52.4**	48.0	66.9	94.5
ODIN	85.3	92.4	99.4	27.8	68.2	56.8	26.8	60.2	54.5	64.0	61.5	**52.4**	45.7	66.6	94.4
Mahalanobis	91.7	**96.4**	99.8	71.1	94.7	67.1	67.3	96.1	64.2	**100.0**	**98.2**	**52.4**	96.4	56.6	**100.0**
MC Dropout	86.3	93.4	99.5	32.0	74.1	58.9	29.5	67.0	56.1	73.3	70.8	**52.4**	51.8	64.9	94.9
Feature Squeezing	80.5	96.2	99.4	64.4	97.9	66.7	67.0	**98.1**	64.2	99.9	96.6	**52.4**	95.6	66.5	**100.0**
Learned Deferral	**98.3**	83.0	**100.0**	68.3	94.1	66.3	66.9	93.0	63.8	99.9	94.6	**52.4**	**97.1**	**73.9**	99.9

The sweep shows a trade-off the attacker cannot tune away. Raising λ does reduce disagreement's ability to detect the attack (AUROC 77 $\rightarrow$ 61), but the budget diverted to agreement is budget not spent on fooling the models: the teacher's accuracy under attack rises from 17% to 52% and the students' from 58% to 81%. Disagreement's error-ranking AUROC is flat at 82–84 at every λ, so whatever the attacker chooses, the framework still knows which teacher predictions to trust, and at a 50% budget it recovers 65–89% oracle accuracy against 51–77% for confidence. The setting most damaging to the framework is $\lambda \approx 1$, where the teacher is most broken and detection is still 73; that is the setting reported in the main text. Confidence remains inverted (19–51) at every λ, and Mahalanobis, which the objective does not target, holds its detection AUROC at 76–82 while never ranking teacher errors above 70.

Table 11: MNIST at $\epsilon = 0.2$. Columns as in Table 8. Teacher accuracy: FGSM 17.8, PGD 0.0, C&W 0.0, AA 0.0, Adapt 0.2; student majority 98.3 / 98.8 / 98.8 / 98.9 / 95.5. Err is undefined (,) where the teacher has no correct predictions.

	FGSM			PGD			C&W			AutoAttack			Adaptive		
	Err	Adv	Or@50	Err	Adv	Or@50	Err	Adv	Or@50	Err	Adv	Or@50	Err	Adv	Or@50
S1 Low Confidence	65.3	99.1	61.7	—	20.6	**50.0**	—	13.9	**50.0**	—	22.9	**50.0**	0.8	26.2	50.0
S2 Disagreement	**99.2**	90.7	**67.7**	—	99.3	**50.0**	—	99.3	**50.0**	—	99.4	**50.0**	99.1	98.7	**50.2**
S3 S1 and S2	99.1	90.8	**67.7**	—	99.0	**50.0**	—	99.0	**50.0**	—	99.1	**50.0**	99.1	98.4	**50.2**
S4 S1 or S2	**99.2**	99.4	**67.7**	—	99.1	**50.0**	—	99.1	**50.0**	—	99.5	**50.0**	98.3	98.4	**50.2**
S5 Predictive Entropy	50.8	36.3	58.7	—	41.6	**50.0**	—	42.0	**50.0**	—	39.7	**50.0**	55.4	74.5	50.1
S6 Vote Margin	50.4	50.1	59.5	—	50.2	**50.0**	—	50.3	**50.0**	—	50.1	**50.0**	49.6	52.3	**50.2**
S7 Mutual Information	35.0	56.6	55.5	—	56.7	**50.0**	—	56.9	**50.0**	—	56.9	**50.0**	40.0	65.4	50.0
Max-Logit	65.6	98.4	61.7	—	12.2	**50.0**	—	7.1	**50.0**	—	18.0	**50.0**	0.7	17.5	50.0
Teacher Entropy	66.3	99.0	61.7	—	18.8	**50.0**	—	12.0	**50.0**	—	21.6	**50.0**	0.7	24.3	50.0
Energy	66.6	98.2	61.8	—	11.2	**50.0**	—	6.6	**50.0**	—	17.3	**50.0**	0.7	16.4	50.0
ODIN	64.8	98.5	61.6	—	9.0	**50.0**	—	5.8	**50.0**	—	16.5	**50.0**	0.7	14.4	50.0
Mahalanobis	68.2	**100.0**	62.8	—	**99.6**	**50.0**	—	**99.5**	**50.0**	—	**99.8**	**50.0**	30.5	**99.5**	50.1
MC Dropout	66.3	99.0	61.8	—	19.1	**50.0**	—	12.5	**50.0**	—	21.8	**50.0**	0.7	24.7	50.0
Feature Squeezing	68.5	98.5	62.3	—	98.0	**50.0**	—	98.7	**50.0**	—	97.8	**50.0**	14.8	97.9	50.0
Learned Deferral	86.9	97.8	66.3	—	60.8	**50.0**	—	73.2	**50.0**	—	74.5	**50.0**	7.6	74.2	50.0

Table 12: Tiny ImageNet at $\epsilon = 2/255$. Columns as in Table 8. Teacher accuracy: FGSM 22.7, PGD 8.2, C&W 8.6, AA 3.5, Adapt 10.4; student majority 26.3 / 26.4 / 26.4 / 26.4 / 24.8. With students 44 points behind the teacher on clean data, disagreement remains well below feature squeezing on error ranking at every budget.

	FGSM			PGD			C&W			AutoAttack			Adaptive		
	Err	Adv	Or@50	Err	Adv	Or@50	Err	Adv	Or@50	Err	Adv	Or@50	Err	Adv	Or@50
S1 Low Confidence	57.8	55.2	64.3	18.1	29.5	50.5	20.6	28.0	51.0	75.8	42.4	53.3	28.2	29.8	52.2
S2 Disagreement	71.9	53.9	66.4	78.3	57.2	56.6	76.6	57.1	56.7	81.2	58.5	53.0	76.7	56.2	58.2
S3 S1 and S2	73.9	54.3	66.6	67.6	40.2	55.4	65.2	39.2	55.5	88.4	50.2	53.4	67.7	40.0	56.8
S4 S1 or S2	71.8	54.9	66.2	76.3	57.3	56.6	74.9	57.1	56.6	82.8	58.6	53.0	75.2	56.2	58.3
S5 Predictive Entropy	65.4	50.1	64.8	73.3	50.1	56.3	72.5	50.1	56.6	79.0	50.1	53.0	70.9	50.6	57.7
S6 Vote Margin	59.3	50.0	64.1	64.3	50.0	55.8	62.0	50.0	55.7	65.0	50.1	52.6	62.7	50.3	57.0
S7 Mutual Information	35.0	49.9	57.9	26.8	49.9	51.9	27.0	49.9	52.0	19.9	49.9	50.4	29.5	49.4	52.7
Max-Logit	49.1	56.7	61.5	12.1	28.1	50.2	18.8	25.3	50.8	64.9	40.6	52.8	19.4	28.6	50.9
Teacher Entropy	54.1	56.1	62.9	15.2	29.1	50.3	21.0	26.9	51.0	73.1	41.7	53.3	24.9	29.4	51.8
Energy	44.0	57.8	59.9	11.0	28.9	50.1	19.8	25.5	50.8	62.4	40.4	52.6	17.8	29.3	50.8
ODIN	37.5	58.9	57.6	10.6	30.1	50.2	18.6	26.6	50.7	52.7	39.9	52.0	16.2	30.7	50.4
Mahalanobis	76.6	50.5	68.1	69.8	46.9	56.1	70.2	48.3	56.4	**97.1**	53.9	**53.5**	77.6	45.8	58.5
MC Dropout	54.2	56.1	63.0	15.3	29.2	50.3	21.1	27.1	51.0	73.5	41.7	53.3	25.2	29.6	51.8
Feature Squeezing	**83.8**	**61.3**	**69.7**	**89.2**	**76.6**	**58.1**	**87.7**	**75.9**	**58.2**	95.8	**76.7**	53.4	**90.2**	**74.6**	**60.2**
Learned Deferral	70.8	53.1	65.6	51.0	40.6	54.1	47.0	40.5	53.9	94.3	49.1	**53.5**	63.1	39.7	56.5

Table 13: Tiny ImageNet at $\epsilon = 4/255$. Columns as in Table 8. Teacher accuracy: FGSM 14.9, PGD 0.8, C&W 0.5, AA 0.1, Adapt 2.5; student majority 26.1 / 26.1 / 26.2 / 26.1 / 21.9. Err values in grey rest on ≤ 65 surviving teacher predictions.

	FGSM			PGD			C&W			AutoAttack			Adaptive		
	Err	Adv	Or@50	Err	Adv	Or@50	Err	Adv	Or@50	Err	Adv	Or@50	Err	Adv	Or@50
S1 Low Confidence	47.7	55.5	57.5	4.4	16.9	50.0	4.9	12.6	50.0	55.1	35.7	**50.1**	10.2	17.9	50.0
S2 Disagreement	74.7	55.6	61.3	85.3	59.5	50.7	85.1	59.5	50.4	87.9	59.9	**50.1**	80.9	57.8	52.2
S3 S1 and S2	72.9	55.7	60.5	76.6	34.7	50.6	77.0	32.4	50.4	88.5	47.2	**50.1**	70.7	34.4	51.8
S4 S1 or S2	73.8	56.5	61.2	82.4	59.1	50.7	82.1	58.9	50.4	88.7	60.0	**50.1**	77.8	57.2	52.2
S5 Predictive Entropy	67.3	50.2	60.1	85.1	50.4	**50.8**	85.3	50.4	50.4	98.1	50.3	**50.1**	77.0	51.2	52.1
S6 Vote Margin	61.5	50.1	59.7	67.5	49.9	50.6	69.1	50.0	50.3	75.1	50.1	**50.1**	65.7	50.6	51.8
S7 Mutual Information	33.3	49.8	54.7	13.6	49.6	50.0	12.5	49.6	50.0	0.9	49.7	50.0	23.2	48.8	50.4
Max-Logit	37.7	58.2	54.8	2.3	14.7	50.0	4.1	10.8	50.0	42.9	33.6	50.0	5.5	16.1	50.0
Teacher Entropy	41.7	57.3	56.0	3.3	16.3	50.0	4.9	12.1	50.0	51.1	35.1	**50.1**	8.1	17.5	50.0
Energy	31.0	60.2	53.7	2.0	15.3	50.0	4.7	11.2	50.0	40.6	33.3	50.0	4.9	16.8	50.0
ODIN	26.2	**61.3**	52.4	1.1	13.0	50.0	3.3	10.9	50.0	28.7	29.8	50.0	3.1	15.4	50.0
Mahalanobis	70.9	49.8	61.6	75.4	56.3	50.7	89.5	62.7	**50.5**	100.0	58.8	**50.1**	80.7	53.3	52.1
MC Dropout	41.8	57.3	56.1	3.4	16.5	50.0	5.4	12.3	50.0	52.1	35.2	**50.1**	8.4	17.7	50.0
Feature Squeezing	**80.1**	61.2	**63.1**	86.6	**80.9**	**50.8**	88.6	**81.4**	**50.5**	99.2	**79.2**	**50.1**	**88.4**	**79.1**	**52.5**
Learned Deferral	60.6	53.7	58.7	37.7	41.1	50.3	53.3	41.1	50.3	99.9	50.7	**50.1**	54.4	40.3	51.4

Table 14: GTSRB under the ensemble-aware adaptive attack at four values of λ. Err = error-detection AUROC; Adv = adversarial-detection AUROC; Or@50 = oracle accuracy at a 50% query budget. Teacher / student-majority accuracy under attack is given in the header. Best per column in **bold**.

Teacher / students acc. (%)	$\lambda = 0.3$ 17.3 / 57.8			$\lambda = 1$ (main text) 15.3 / 60.7			$\lambda = 3$ 33.7 / 76.1			$\lambda = 10$ 51.9 / 81.2		
	Err	Adv	Or@50	Err	Adv	Or@50	Err	Adv	Or@50	Err	Adv	Or@50
S1 Low Confidence	24.2	45.1	52.8	20.8	49.0	51.0	40.3	59.8	63.9	51.0	62.8	76.7
S2 Disagreement	81.6	77.2	65.4	83.7	73.2	65.2	84.3	67.8	**81.5**	82.8	60.7	88.6
S3 S1 and S2	79.8	75.3	65.0	83.6	71.9	65.2	84.6	67.3	81.3	83.1	60.6	87.9
S4 S1 or S2	73.3	75.9	**65.6**	68.9	70.7	**65.2**	73.8	70.6	80.2	74.0	66.9	85.3
S5 Predictive Entropy	70.4	63.9	61.7	**88.0**	67.1	64.5	**85.2**	65.4	79.2	**88.7**	62.4	**91.5**
S6 Vote Margin	50.3	55.9	58.8	62.8	56.4	59.8	62.5	53.3	70.7	62.4	50.8	79.9
S7 Mutual Information	58.9	64.0	60.4	84.1	66.1	63.8	83.1	63.5	78.4	86.9	59.6	90.9
Max-Logit	22.2	39.6	52.2	19.2	43.8	50.9	36.8	55.6	62.7	46.9	60.0	75.1
Teacher Entropy	24.0	44.4	52.8	20.5	48.3	51.0	39.9	59.2	63.8	50.6	62.4	76.5
Energy	22.2	39.6	52.2	19.0	43.7	50.9	36.6	55.6	62.7	46.7	59.9	75.1
ODIN	21.9	41.1	52.0	18.5	45.2	50.7	35.6	56.7	62.3	44.9	60.8	74.1
Mahalanobis	59.2	**81.1**	60.2	49.6	**82.3**	57.3	64.6	**80.0**	71.8	70.2	**75.7**	83.4
MC Dropout	26.9	45.8	53.8	22.6	49.4	51.7	42.5	59.0	64.8	52.5	61.3	77.3
Feature Squeezing	42.2	67.3	56.3	40.1	67.3	55.1	57.9	70.2	69.3	64.4	67.7	81.1
Learned Deferral	41.3	59.4	56.5	42.2	60.3	55.7	61.0	66.0	70.9	68.3	64.6	83.4