Received XX Month, XXXX; revised XX Month, XXXX; accepted XX Month, XXXX; Date of publication XX Month, XXXX; date of current version XX Month, XXXX.

Digital Object Identifier 10.1109/XXXX.2022.1234567

ADS-C: Antidistillation Sampling for Classification

Khawaja Abaid Ullah[1], Mohammad Javad Khojasteh[1]

[1]Department of Electrical and Microelectronic Engineering, Rochester Institute of Technology, Rochester, NY 14623 USA

Corresponding author: Khawaja Abaid Ullah (email: ka4150@rit.edu).

ABSTRACT Knowledge distillation enables an adversary to replicate a proprietary classifier by querying its prediction interface and training a surrogate on the returned probability vectors. Antidistillation sampling, proposed for large language models, counters this threat with an input-dependent, gradient-directed perturbation of the served distribution; its transfer to classification has not been studied. Adapting the defense to classification, we show its behavior is governed by the distribution of the teacher's per-input confidence margins. Because well-trained classifiers are severely overconfident, the direct transfer exhibits an inert window: below a closed-form-predictable threshold, it affects neither attacker nor defender; beyond it, the defense undergoes a phase transition and degrades the teacher faster than the attacker's student. Temperature softening rescales the transition in closed form, and every temperature configuration lies on the same unfavorable trade-off curve. Our method, ADS-C, composes the perturbation under a closed-form, per-input margin budget that provably preserves every served top-1 prediction, so the defended teacher's accuracy equals the undefended teacher's identically. Under this guarantee the distilled student still loses 17.4 percentage points on CIFAR-100, 29.6 on CIFAR-10, and 13.3 on Tiny-ImageNet; matching this degradation with the unmodified defense costs 27.5, 32.9, and 22.2 points of teacher accuracy. Because served labels are unchanged, a hard-label attacker gains nothing, while the defended soft output trains a student up to 29.7 points below that floor: the incentive to distill served probabilities is not merely removed but reversed. To our knowledge, ADS-C is the first antidistillation defense for classification whose utility cost is exactly zero.

INDEX TERMS Antidistillation, knowledge distillation, model extraction, inference-time defenses, machine learning security.

I. INTRODUCTION

KNOWLEDGE distillation (KD) [1], [2] transfers the predictive function of a large *teacher* model into a smaller *student* model by training the student on the teacher's returned probability vector outputs. This very same mechanism that makes possible the compression of large models into smaller, efficient networks [3]–[6], also enables *model extraction* wherein an adversary with black-box Application Programming Interface (API) access systematically queries a deployed classifier, collects the returned probability vectors, and trains a student model that replicates the teacher's behavior at a fraction of the original cost [7], [8]. As recent events have highlighted, this threat is not just hypothetical. DeepSeek's 2025 release demonstrated that near-frontier capability may be obtained largely through distillation from proprietary models [9]. Additionally, Anthropic recently disclosed an industrial-scale extraction campaign against Claude that produced over 16 million exchanges via approximately 24,000 fraudulent accounts [10]. Beyond direct intellectual-property loss [11], [12], extracted models facilitate downstream membership-inference [13], [14], evasion [15]–[17], and privacy attacks against the underlying training distribution [18]–[20], raising the stakes especially in regulated domains such as medicine, and finance where machine learning applications are gaining traction [21], [22]. Beyond these settings, the implications of model extraction also extend to Cyber-Physical Systems (CPS) as KD is increasingly used in resource-constrained robotic platforms [23]. Distillation-based extraction may also expose safety-critical control behavior in the physical world, raising security concerns for CPS deployments.

The attacker consumes the returned probability vector, also known as the *soft labels*, in contrast to *hard labels*, where the output contains only the index of the predicted class. The inter-class structure of these soft labels, i.e., the relative probabilities of the classes other than the top class, constitutes the

dark knowledge of the given model [2]. This dark knowledge makes soft-label distillation markedly more effective than training on hard labels alone. A naive defense against such distillation attempts would be to always return uninformative output, but that would make the deployed model useless for regular users. The natural question is therefore how to corrupt the dark knowledge a model deployed through an API serves without destroying the utility of the given model for legitimate users. Every inference-time defense implicitly negotiates this trade-off, and we argue the trade-off rate should be measured against an explicit break-even: a defense that removes one percentage point (pp) of attacker accuracy per pp of teacher accuracy, a 1:1 trade-off, is no better than deploying a worse model. A useful defense must beat the 1:1 trade-off, ideally by a wide margin in favor of the model being defended.

Antidistillation Sampling (ADS) [24] is a compelling starting point. Proposed for autoregressive language models, ADS perturbs each next-token distribution with an *input-dependent, gradient-directed* penalty, computed from a hidden proxy student, that estimates how much serving each token would help a distilling student. Unlike static perturbation defenses, which apply a fixed transformation that an attacker can invert or meta-learn [25], [26], the ADS perturbation is recomputed per query from hidden state. Whether this idea transfers to classification, which is among the most broadly exposed extraction surfaces in practice, has remained open. The question is not merely one of engineering: in a classification API the distilling student ingests the *entire* served vector through its KL objective, whereas LLM sampling collapses the perturbation onto a single emitted token, so per unit of perturbation ADS should be *more* potent in classification, not less.

This paper adapts ADS to classification, analyzes the mechanism that governs its behavior there, and repairs its ineffectiveness with a composition rule whose utility cost is provably zero. Our contributions are as follows.

1) **Adaptation.** We instantiate the ADS penalty in the classification setting, preserving its derivation, and identify the structural differences from the LLM setting (Sec. IV).
2) **Diagnosis.** We show that the behavior of the transferred defense is determined by the distribution of per-input *flip thresholds*, whose cumulative distribution, before any attack or defense is run, accurately predicts the defended teacher's accuracy cost (Sec. V). Classifier overconfidence [27] renders the defense inert at low strengths and unfavorable at high strengths; the perturbation itself is faithful and correctly aimed at dark knowledge, but *margin-blind*.
3) **Analysis of temperature softening.** We show that softening the served distribution rescales the phase transition in closed form, and confirm the prediction in trained students: softening amplifies the perturbation at low strengths while collapsing the teacher at moderate ones, and no temperature configuration improves the trade-off (Sec. VI).
4) **ADS-C.** We compose the ADS perturbation under a closed-form, per-input, margin-aware budget: each input receives the largest penalty strength that provably keeps its served top-1 prediction unchanged (Proposition 1), so the defended teacher's accuracy equals the undefended teacher's identically. The surviving perturbation still degrades the distilled student substantially, by $17.4\,\mathrm{pp}$ on CIFAR-100, $29.6\,\mathrm{pp}$ on CIFAR-10, and $13.3\,\mathrm{pp}$ on Tiny-ImageNet at zero utility cost (Secs. VII–VIII). Against the complementary hard-label attacker the defended soft output proves strictly worse as a distillation signal than the argmax it accompanies. The soft-label student falls $17.1\,\mathrm{pp}$ and $29.7\,\mathrm{pp}$ below what the hard-label attacker achieves through distillation (Sec. VIII-E).
5) **Relaxations of the guarantee.** We introduce two mechanisms with closed-form-predictable utility cost: first, stochastic enforcement of the budget; second, a minimum served strength. Both trade a controlled amount of the guarantee for attacker uncertainty. We price each relaxation against the attacker's best response between soft- and hard-label extraction, and report a negative result on margin-ranked sacrifice that further clarifies the damage mechanism (Sec. IX).
6) **Evaluation methodology and reproducibility.** We argue that once utility cost is zero, defenses should be compared at matched served fidelity rather than matched penalty strength, and show that the two comparisons can disagree (Sec. VIII-D). We release a library and every experiment driver: https://github.com/the-kas-lab/ADS-C.

II. RELATED WORK

Distillation and dark knowledge. Distillation trains a student on a teacher's output distributions [1], [2], [28], [29]; its advantage over hard-label training is attributed to the inter-class structure of those distributions [2], [30], [31]. Furlanello *et al.* [30] showed diagnostically that destroying non-target structure destroys most of distillation's benefit, providing evidence that dark knowledge is the asset a defense should target.

Model extraction. Extraction attacks replicate a deployed model from query access [7], [32]–[34]; surveys [8], [35]–[37] document their practicality. We consider the distillation-based extraction attacker, who trains on returned soft labels which is the strongest and most common variant when probability vectors are exposed.

Extraction defenses. Training-time defenses harden the teacher's weights [38]–[41] but require retraining and access to the original corpus, which is often undesirable or even infeasible at deployment scale. Detection and monitoring [42]–[44] flag anomalous query streams, but rely on out-of-distribution assumptions that prove futile against attackers

with in-distribution data [41] and are circumvented by Sybil accounts [45], [46]. Watermarking [47]–[50] supports post-hoc attribution but erects no barrier against the extraction itself. *Perturbation* defenses are closest to our approach. These defenses modify the served probabilities: Reverse Sigmoid [25], prediction poisoning [51], ModelGuard [52], noise-transition learning [53], and information-theoretic output compression [54]. All of these pay for protection with utility, and the static members of the family being fixed functions of the input are invertible by meta-learned recovery attacks such as D-DAE [26]. None offers a per-query *guarantee* on the served top-1 prediction.

ADS. Antidistillation Sampling [24] is input-dependent and gradient-directed wherein a hidden proxy student defines, per query, a direction in output space that maximally harms a distilling student. It was derived and evaluated for LLM reasoning traces. Two questions are open: whether it transfers to classification, and what it costs the defender there. We answer both: the transfer's behavior is gated entirely by classifier overconfidence, and the repair (a margin budget at composition) makes the cost exactly zero.

III. THREAT MODEL AND PROBLEM SETUP

Figure 1 summarizes the setting. A defender trains a classifier $\mathcal{T}\colon \mathcal{X} \to \Delta^{K-1}$ mapping an input space $\mathcal{X}$ to the probability simplex over K classes and deploys it behind a query API that returns, for each query $\mathbf{x} \in \mathcal{X}$, a served probability vector $\hat{\mathbf{p}}(\mathbf{x}) \in \Delta^{K-1}$.

Adversary. The adversary holds (i) black-box access to the API with an arbitrary but bounded query budget; (ii) an unlabeled query corpus drawn from the same distribution $\mathcal{D}$ over $\mathcal{X}$ as the teacher's training data—we make no out-of-distribution assumption, which disables detection-style defenses by construction; and (iii) a student architecture $\mathcal{S}$ of its choosing. The adversary lacks the teacher's parameters, and ground-truth labels. With no labels, the attacker's objective reduces from the standard hard-plus-soft KD loss to pure soft-label matching,

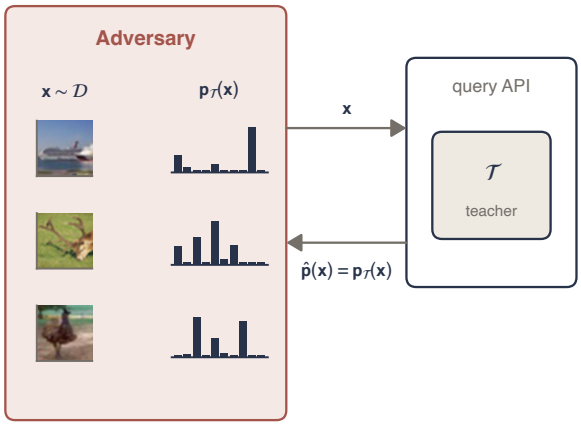
$$\mathcal{L}_{\mathrm{UKD}} \;=\; D_{KL}\big(\hat{\mathbf{p}}(\mathbf{x}) \,\big\|\, \mathbf{p}_{\mathcal{S}}(\mathbf{x})\big), \tag{1}$$

evaluated at temperature 1: the served vector is the attacker's sole supervisory signal, and hence the natural locus of defense.

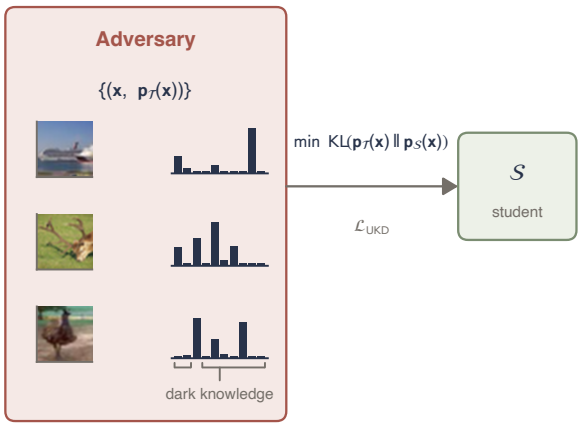
Why the soft-label attacker. We scope the threat to the *dark-knowledge distiller*, and this scoping is principled rather than convenient. The attacker's advantage over training on public hard labels is the served soft-label structure; an attacker who reads only the served argmax gains nothing that a labeling function of equal top-1 accuracy would not provide. Our defense will, by construction, never alter the served argmax (Sec. VII), so hard-label distillation against the defended API yields exactly what it would yield against the undefended API (verified empirically in Sec. VIII-E): the defense removes the soft-label advantage. The exploitation of the guarantee's determinism itself is treated in Sec. IX.

Defender. The defender owns the teacher, a modest labeled holdout set, and a compact *proxy* student $\mathcal{P}$ that stands in for the unknown attacker student; it cannot retrain the teacher and must act at inference time only. The defender's goals are: (a) *utility*: legitimate users, who consume the top-1 prediction, must be unharmed; (b) *distillation resistance*: a student trained via (1) on the served outputs should be substantially degraded.

Metrics. (A summary of notation is provided in Appendix XI-A.) We report three quantities. *Teacher top-1*: accuracy of the served argmax. *Student top-1*: accuracy of the distilled student. *Served fidelity*: $D_{KL}(\mathbf{p}_{\mathcal{T}} \,\|\, \hat{\mathbf{p}})$ averaged over the evaluation set, the total distributional perturbation the defense spends; this axis quantifies the defense's "corruption budget" and becomes essential once utility cost is zero (Sec. VIII-D).



(a) The adversary constructs a labeled dataset by querying the victim API with unlabeled samples.



(b) The adversary trains a surrogate student on the returned soft labels.

FIGURE 1. The unauthorized knowledge-distillation threat model.

IV. ANTIDISTILLATION SAMPLING FOR CLASSIFICATION

A. Preliminaries

A pretrained teacher with parameters $\theta_{\mathcal{T}}$ produces logits $\mathbf{z}_{\mathcal{T}} = \mathcal{T}(\mathbf{x};\theta_{\mathcal{T}}) \in \mathbb{R}^K$; the softmax $\sigma(\mathbf{z})_i = \exp(z_i)/\sum_j \exp(z_j)$ yields probabilities $\mathbf{p}_{\mathcal{T}}$. Only logit differences matter: adding a constant to $\mathbf{z}$ leaves $\sigma(\mathbf{z})$ unchanged, so the margin $z_t - z_j$ between the top class t and a rival j is the invariant quantity that decides the argmax. Distillation trains a student to minimize (1) against the served distribution.

B. The ADS penalty

ADS asks, per query and per class: *if the student came to believe class y is slightly more likely on this input, would that make the student worse?* Formally, let $\ell(\theta_{\mathcal{P}})$ be a downstream loss of the proxy on the defender's labeled holdout. We take $\ell(\theta_{\mathcal{P}}) = \frac{1}{N}\sum_{i=1}^{N} D_{KL}\big(\mathbf{p}_{\mathcal{T}}(\mathbf{x}_i) \,\|\, \mathbf{p}_{\mathcal{P}}(\mathbf{x}_i)\big)$, the proxy's mean disagreement with the teacher over the holdout, following [24], and let $g = \nabla\ell(\theta_{\mathcal{P}})$ be its gradient. One SGD step of distillation on class y moves $\theta_{\mathcal{P}}$ along $\nabla_\theta \log p(y \mid \mathbf{x};\theta_{\mathcal{P}})$; the resulting change in ℓ is, to first order, the inner product of that update with $\nabla\ell$. Exactly as in the LLM derivation, symmetry of the directional derivative converts this into a finite difference over two *frozen clones* of the proxy, $\theta_{\mathcal{P}\pm} = \theta_{\mathcal{P}} \pm \varepsilon g$:

$$\widehat{\Delta}(y \mid \mathbf{x}) \;=\; \frac{\log p(y \mid \mathbf{x};\theta_{\mathcal{P}+}) - \log p(y \mid \mathbf{x};\theta_{\mathcal{P}-})}{2\varepsilon}. \qquad (2)$$

A positive $\widehat{\Delta}(y \mid \mathbf{x})$ means that serving mass on class y raises the proxy's holdout loss, which is good for the defender. The defended distribution adds the penalty to the teacher's log-probabilities:

$$\hat{\mathbf{p}}(\cdot \mid \mathbf{x}) \;=\; \sigma\big(\log \mathbf{p}_{\mathcal{T}}(\cdot \mid \mathbf{x}) + \lambda\, \widehat{\Delta}(\cdot \mid \mathbf{x})\big), \qquad (3)$$

with $\lambda \geq 0$ the defense strength; $\lambda = 0$ recovers the undefended API.

Parameter choices. (i) Defining ℓ as a per-sample mean fixes the scale of g independently of the holdout size; empirically $\|g\|$ is of order one (1.00 on CIFAR-100, 0.77 on CIFAR-10, 1.37 on Tiny-ImageNet), so λ operates on a comparable scale across datasets. We sweep $\lambda \in [0, 3]$, extended to $\lambda = 10$ for unmodified ADS to map the saturation regime (Sec. V-D). (ii) The step ε is calibrated, not tuned: we select ε to minimize the discrepancy between (2) and the exact Jacobian–vector product it approximates, obtaining $\varepsilon = 3 \times 10^{-3}$ on both CIFAR datasets and $\varepsilon = 10^{-2}$ on Tiny-ImageNet, each with a magnitude ratio of ≈ 1.00 (Appendix XI-B); the finite difference is a faithful estimate, which will matter for the diagnosis of Sec. V. (iii) The proxy (ResNet-20 on the CIFAR datasets, ResNet-18 on Tiny-ImageNet) is substantially smaller than the teacher and is trained on the teacher's training split, with a disjoint 10% held out for computing g; the two clones are built once, offline, and frozen, so a defended query costs two small forward passes and no gradient computation. (iv) The margin floor m introduced in Sec. VII is analyzed in Appendix XI-C.

C. ADS Case for Classification

The penalty (2) is structurally identical to the next-token formulation of [24]; what differs is downstream consumption. In autoregressive generation the perturbation tilts the next-token distribution, but the attacker observes only a single sampled token per step: a one-hot realization whose expectation carries the tilted distribution, so the signal on the non-sampled coordinates of $\widehat{\Delta}$ reaches the student only in expectation, across many draws. In classification distillation, the attacker observes the served vector itself: the KL objective (1) ingests every coordinate of $\widehat{\Delta}$ exactly, on every query. The same perturbation is thus delivered through a noiseless channel rather than estimated through a stochastic one, and per unit of perturbation ADS should be *more* potent in classification, not less. The next section shows that, despite this, the direct transfer is inert over a wide regime, then trades at an unfavorable rate, and finally saturates with both parties near their accuracy floors. In addition, it locates the cause of this behavior not in the penalty but in its composition with the teacher.

V. DIAGNOSIS OF ADS INEFFECTIVENESS

Swept over λ, the unmodified ADS exhibits three behaviors in sequence that make it undesirable for deployment in the classification domain. For $\lambda \lesssim 0.2$ it is *inert*: on both CIFAR datasets, student and teacher accuracy each change by at most 1–2 pp, within seed noise. The defense measurably affects neither party. Beyond that, it becomes active at an unfavorable rate: on CIFAR-100 at $\lambda = 0.5$, the distilled student drops 5.9 pp while the teacher drops 12.2 pp, a trade-off rate of 0.48, half of break-even; on CIFAR-10 the rate is 0.85 (9.6 pp student for 11.3 pp teacher). Finally, for $\lambda \gtrsim 2$ the sweep *saturates*: the trade-off remains sublinear while both models degrade toward their floors, and by $\lambda = 3$ the defended teacher's accuracy has fallen to the level of the very student it defends against (Fig. 9, diagonal curves). This section explains all three behaviors through a single mechanism and derives the design requirement for the remedy.

A. Overconfidence of Classifiers

Modern, well-trained networks are systematically overconfident [27], [55]: their top-class probability $p_{\max}$ concentrates near 1. Table 1 quantifies this for our teachers. On CIFAR-10 the median served $p_{\max}$ exceeds 0.99999. The median non-target mass, the entire surface on which dark knowledge lives and on which an additive penalty can visibly act, is below 10^{-6}. CIFAR-100 is less extreme (median non-target mass $\approx 6 \times 10^{-3}$), and Tiny-ImageNet sits between the two (4×10^{-4}) despite having twice CIFAR-100's classes. The comparison is deliberate: the Tiny-ImageNet teacher is a strong ImageNet-pretrained ResNet-50 fine-tune, whereas

TABLE 1. Teacher confidence structure (median over the evaluation set).

Benchmark	K	$p_{\max}$	Non-target mass
CIFAR-10 (RN-56)	10	> 0.99999	8×10^{-7}
Tiny-ImageNet (RN-50†)	200	0.9996	4×10^{-4}
CIFAR-100 (RN-110)	100	0.994	6×10^{-3}

†ImageNet-pretrained, fine-tuned at 64×64.

the CIFAR teachers are trained from scratch, and the pretrained model's strength overrides the intuition that more classes produce more diffuse outputs. We find that overconfidence is determined by model quality, not class count, and strong pretrained models are precisely what production APIs deploy, so the overconfident regime studied here is the realistic one rather than a small-benchmark artifact. (At the extreme of output dimension, LLMs over >50k-token vocabularies operate in a naturally diffuse regime which is one reason the original ADS formulation could presuppose sufficient non-target mass.)

The overconfidence is also *heterogeneous*, and this is the central observation. The top-logit margin $z_t - z_{(2)}$ (top minus runner-up) spans two orders of magnitude within a single dataset: on CIFAR-100 its 10th/50th/90th percentiles are 0.7/5.5/15.0 log-units; on Tiny-ImageNet, 1.1/8.3/17.0; on CIFAR-10, 3.9/14.4/21.0. Most inputs have margins far too large for a moderate output perturbation to overcome, while a persistent minority — 13.4%, 9.4%, and 2.4% of inputs, respectively, have margins below one log-unit—lie close to a prediction flip. Any defense that treats all inputs identically will therefore be simultaneously too weak on the head of this distribution and too strong on its tail. To our knowledge, no prior antidistillation work has examined how this confidence structure interacts with an output-perturbing defense; the remainder of this section shows the interaction is decisive.

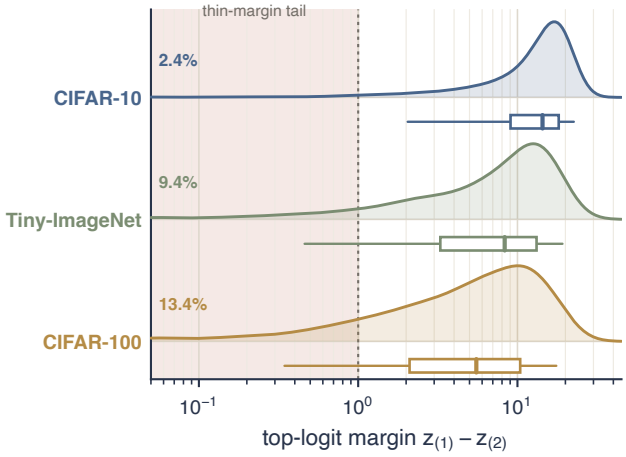


FIGURE 2. Overconfidence and its heterogeneity. Per-dataset distribution of the top-logit margin $z_{(1)} - z_{(2)}$ (top class minus runner-up), drawn as a density (above) and a box (below) that captures the median, interquartile range, and p5–p95 whisker on a logarithmic axis as within every dataset the margin varies over orders of magnitude.

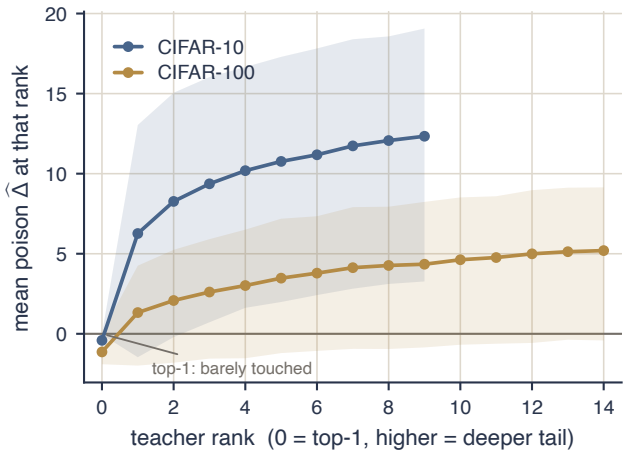


FIGURE 3. Anatomy of the ADS penalty. Mean $\widehat{\Delta}$ as a function of the teacher's class rank; near zero at rank 0, positive across the non-target tail. The penalty is aimed at dark knowledge, not at the argmax.

B. Soundness of ADS Penalty

Before locating the futility at composition we rule out the penalty. Two verifications (Fig. 3):

Faithfulness. At the calibrated ε, the finite-difference $\widehat{\Delta}$ matches the exact directional derivative with magnitude ratio ≈ 1.00 on both datasets (Appendix XI-B). The poison is not a numerical artifact.

Aim. $\widehat{\Delta}$ concentrates exactly where dark knowledge lives. Averaged over inputs, its value at the teacher's top class is near zero (mean -1.1 on CIFAR-100, -0.4 on CIFAR-10) while deep non-target ranks receive large positive pushes (up to $+5.6$ and $+12.3$ respectively); the class each input's poison boosts hardest sits at teacher rank ≥ 5 for 98% of CIFAR-100 and 63% of CIFAR-10 inputs. ADS corrupts the inter-class structure beneath the prediction.

C. Margin Blindness

Whether the served argmax survives (3) is a per-input race between two numbers. For each rival $j \neq t$, let $c_j = \widehat{\Delta}_j - \widehat{\Delta}_t$ be the rate (per unit λ) at which the poison closes the margin to j; we define the poison's *contrast* on input $\mathbf{x}$ as the largest such rate,

$$c(\mathbf{x}) \;=\; \max_{j \neq t} \; c_j, \tag{4}$$

the rate at which the strongest rival gains on the top class t. The served top-1 flips exactly when λ exceeds the *flip threshold*

$$\lambda^*(\mathbf{x}) \;=\; \min_{j \neq t:\, c_j > 0} \; \frac{z_t - z_j}{c_j}, \tag{5}$$

with $\lambda^*(\mathbf{x}) = +\infty$ when no rival satisfies $c_j > 0$. Two measurements now explain the unfavorable exchange. First, the contrast is large: its median is 17.4 (CIFAR-100) and 16.3 (CIFAR-10) log-units per unit λ—comparable to the 90th-percentile margin and $25\times$ the 10th-percentile margin. Second, and critically, the contrast is *margin-blind*: its Spearman correlation with the margin is $\rho = +0.12$ (CIFAR-100) and $+0.18$ (CIFAR-10). The poison's per-input strength is unrelated to the margin it must respect. Nothing in the objective ℓ ever sees the teacher's margins, so nothing in $\widehat{\Delta}$ respects them.

The consequence follows directly from (5) and the thin tail of Fig. 2: as λ grows, served predictions flip in order of their margins, beginning with the thin-margin inputs. At $\lambda = 0.5$, a global penalty has already flipped 26.1% of CIFAR-100 and 14.8% of CIFAR-10 served predictions (Fig. 4). These flips constitute the entirety of the teacher's cost, and they are an inefficient attack, for a reason the two parties' metrics make precise. The teacher's cost is discontinuous in the perturbation: the moment the top two entries cross, a flip registers as a full served error. The student's supervision is the served vector, which is continuous: a flip just past its threshold crosses at near-parity, so the training signal on that row is almost unchanged. Moreover, low-λ flips land exclusively on thin-margin inputs, whose near-uniform outputs carry little of the inter-class structure a distilling student exploits (Sec. IX-B confirms this independently: corruption concentrated on thin-margin rows barely moves the student). The defender thus pays full price for perturbations the student scarcely notices—at $\lambda = 0.5$ on CIFAR-100 the flip component costs the teacher 12.2 pp while accounting for only 3.1 pp of student damage—whereas the argmax-safe corruption underneath is what a distilling student cannot discount; we quantify this decomposition in Sec. VIII-C.

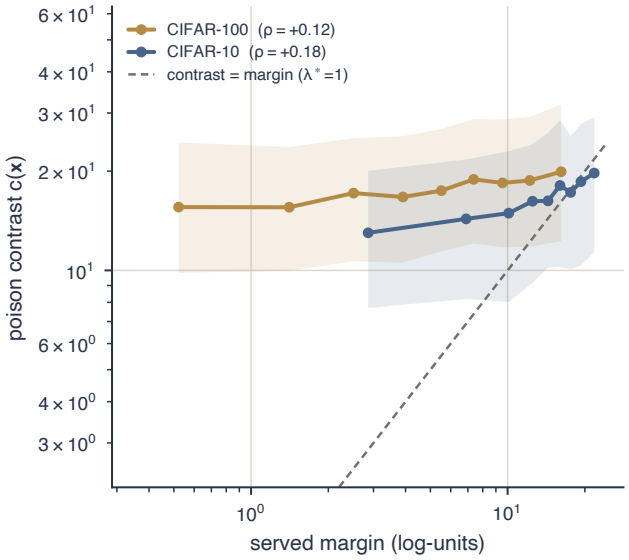
D. The inert window and the phase transition

In addition to explaining the flips, Equations (4)–(5) make the entire teacher-cost curve predictable in closed form, before any student is trained. Let

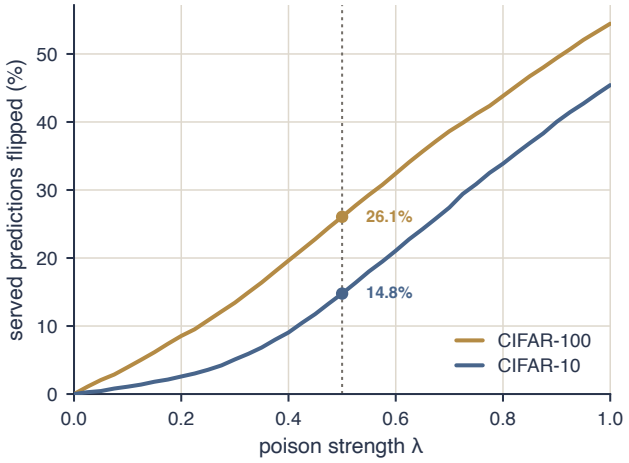
$$F(\lambda) \;=\; \Pr\big[\lambda^*(\mathbf{x}) \le \lambda\big] \tag{6}$$

be the cumulative distribution of the flip threshold over the evaluation set, computable from the teacher's and the proxy's outputs in a single forward pass. The defended teacher's accuracy drop under unmodified ADS is then the fraction of *correctly classified* inputs whose threshold has been crossed. This prediction tracks the measured, trained teacher drop to within ≈1–2 pp at every λ on both datasets (CIFAR-100 at $\lambda = 1$: predicted 34.9 vs. measured 33.3 pp; CIFAR-10 at $\lambda = 2$: 69.8 vs. 69.0 pp). The teacher cost of unmodified ADS is thus fully determined by the flip-threshold distribution.

F inherits its shape from the margin heterogeneity of Sec. V-A, and that shape organizes the sweep into three regimes (Fig. 5):



(a) Margin-blindness



(b) Flip cascade

FIGURE 4. Raw ADS penalty on both teachers. (a) Median poison contrast $c(\mathbf{x})$ against served margin, binned (log–log; band: middle 50%), with the $c(\mathbf{x}) =$ margin boundary at which a served prediction flips at $\lambda = 1$. Each dataset's contrast is *flat* in the margin (Spearman $\rho = +0.12$ and $+0.18$) and sits far *above* the boundary across almost the whole margin range. The push is large (median 17.4/16.3 log-units) and unrelated to the margin it must respect. It meets the boundary only at the largest margins. (b) The consequence of (5): as λ grows, served predictions flip in order of their margins, beginning with the thin-margin tail; by $\lambda = 0.5$ a global penalty has already flipped 26.1% (CIFAR-100) and 14.8% (CIFAR-10) of served predictions.

- **Inert window** (λ below the 5th-percentile flip threshold: $\lambda < 0.12$ on CIFAR-100, $\lambda < 0.30$ on CIFAR-10). $F \approx 0$, and—because the same overconfidence leaves almost no non-target mass for the poison to reweight—the *student* is unaffected as well: measured student and teacher effects are both within 1–2 pp of the undefended anchors, indistinguishable from seed noise. In this regime the defense has no measurable effect on either party.

- **Phase transition** ($0.35 \lesssim \lambda \lesssim 1$). F rises through its steep section; teacher cost and student damage both grow from noise level to tens of pp within about 1.5 octaves of λ—a narrow band of a control parameter that we sweep over four orders of magnitude. We define the *transition threshold* λ_c as the largest λ with $F(\lambda) \leq 1\%$ ($\lambda_c \approx 0.02$ on CIFAR-100, 0.09 on CIFAR-10 by this strict definition; the teacher-cost onset—drop exceeding 1 pp—lies at 0.066 and 0.135, respectively). λ^* is approximately log-normal, so F is a smooth sigmoid in $\log \lambda$ whose inflection lies at the median flip threshold (0.91/1.09).
- **Saturation** ($\lambda \gtrsim 2$). Both parties approach their floors and the sweep exhausts itself. Per-unit-λ student damage falls by two orders of magnitude (20.3 pp near $\lambda = 1$ to 0.2 pp at $\lambda = 10$ on CIFAR-100; 33.6 to 0.3 pp on CIFAR-10) as the student completes its fall to—and slightly through—the chance floor (0.6% vs. 1% chance; 6.3% vs. 10%): deeply poisoned vectors are systematically wrong, not merely uninformative. The teacher's tail is equally exhausted ($64.0 \to 70.2$ pp and $77.6 \to 86.7$ pp of drop over $\lambda = 3 \to 10$), and (6) predicts every measured tail point to within 0.6 pp. By $\lambda = 3$ the defended teacher's accuracy has fallen to the level of the very student it defends against (7.2% vs. 4.2%; 15.1% vs. 12.9%), and by $\lambda = 10$ on CIFAR-10 *below* it (6.0% served vs. 6.3% distilled): the direct transfer does not merely trade unfavorably in this regime—it eliminates, and finally inverts, the accuracy advantage it exists to protect.

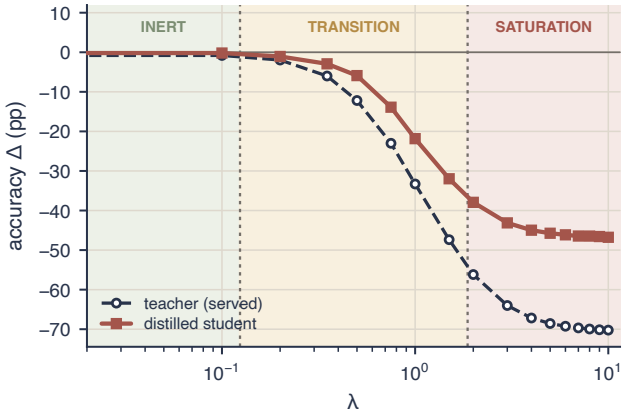
Note the ordering: CIFAR-10, the *more* overconfident teacher, has the *wider* inert window showing that inertness tracks overconfidence, as the mechanism predicts. The same computation places Tiny-ImageNet's transition threshold at $\lambda_c \approx 0.06$, between the two CIFAR values, consistent with its intermediate overconfidence (Table 1); and its closed-form $F(\lambda)$ predicts the measured teacher cost of the unmodified sweep to within 2 pp at every λ (e.g., 33.2 pp predicted versus 31.4 pp measured at $\lambda = 2$), validating the diagnosis on a third teacher whose confidence arises from ImageNet pretraining rather than from a narrow benchmark.

One further property of (6) is used in the next section. Serving the teacher at temperature τ divides every logit gap by τ while leaving the poison contrast unchanged, so every flip threshold is divided by exactly τ. A threshold of λ^*/τ is crossed by a given λ precisely when λ^* is crossed by $\tau\lambda$, so the flipped fraction at strength λ under temperature τ equals the unsoftened fraction at $\tau\lambda$:

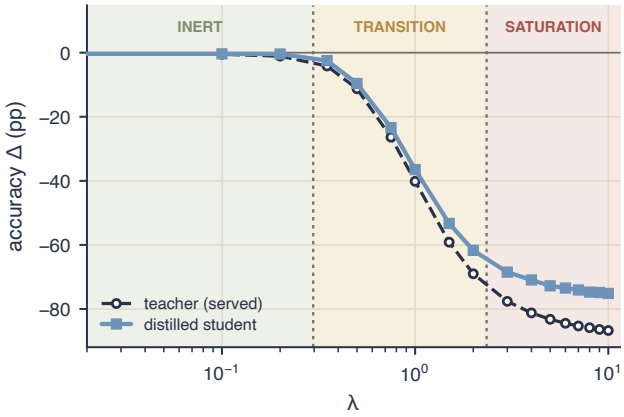
$$F_\tau(\lambda) \;=\; F(\tau\lambda), \tag{7}$$

a parameter-free prediction (Fig. 6): temperature shifts the entire transition curve toward smaller λ by the factor $1/\tau$. Softening does not alter the mechanism; it rescales the λ axis on which the mechanism operates.

The design requirement is now explicit. The penalty direction is sound; what is missing is margin-awareness,



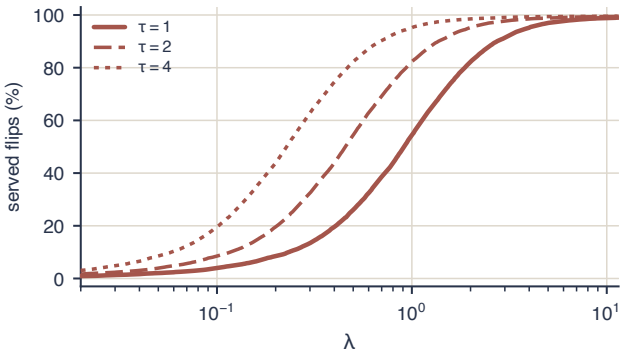
(a) CIFAR-100



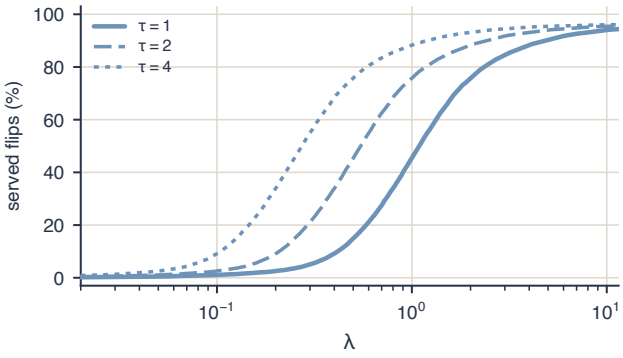
(b) CIFAR-10

FIGURE 5. The inert window and the phase transition (log-λ). Measured accuracy change of *both* parties under unmodified ADS across the three regimes shaded by the flip-threshold CDF (6) (inert: $F \leq 5\%$; saturation: $F \geq 80\%$, near $\lambda \approx 2$). In the inert window ($\lambda < 0.12$ / 0.30) the served teacher and the distilled student are both flat within noise—the defense measurably affects neither party—after which both move together through the transition, the teacher falling at least as fast as the student, into saturation.

and margin-awareness cannot be injected through the poison pipeline. The one free choice inside that pipeline is where the proxy sits, and we tested it: even a proxy distilled directly from the teacher's served probabilities, corresponding to the state a distilling attacker's student would be in during distillation rather than a proxy trained conventionally on labeled data, produces a poison whose contrast–margin correlation remains $\rho \approx 0$ (CIFAR-100). No placement of the proxy makes $\widehat{\Delta}$ margin-aware, because the objective ℓ it derives from never sees a margin. Margin-awareness must therefore be enforced at composition, the only point where the margin and the poison are simultaneously visible. That is ADS-C (Sec. VII). First, however, we examine the natural alternative, whose failure the theory above already predicts and whose failure mode is itself informative.



(a) CIFAR-100



(b) CIFAR-10

FIGURE 6. Temperature slides the transition curve (log-λ). Served flip fraction under temperature τ is $F_\tau(\lambda) = F(\tau\lambda)$ (7): serving at τ divides every flip threshold by τ, shifting the entire transition curve toward smaller λ by the factor $1/\tau$ ($\tau = 2, 4$ slide F left by $2\times, 4\times$). This is a parameter-free prediction—no student is trained—and Secs. VI-B–VI-C confirm it against the trained runs.

VI. TEMPERATURE SOFTENING

Table 1 suggests an obvious remedy, that is, if overconfidence starves ADS of non-target mass, soften the served distribution, with a constant distillation temperature [2], [27] or a confidence-conditioned variant [56]–[58], to create dark knowledge for the poison to corrupt. This intuition is partially correct, but what it yields is a demonstration of the mechanism rather than a viable defense. Softening relaxes both of the constraints that teacher confidence imposes at once. First, it inflates the non-target mass visible to the attacker's KL loss (activating the poison), and secondly, it shrinks every served margin by the same factor (activating the flips, per (7)). We serve $\sigma(\mathbf{z}/T)$ for $T \in \{2, 4\}$, alone and composed with the raw ADS penalty, over $\lambda \in [0, 0.5]$, on both datasets with three seeds, extending the poison-free sweep on CIFAR-10 to $T \in \{6, 8, 10, 12\}$. (Throughout, τ in (7) and T denote the same served temperature; we write T when quoting experimental settings.) We report four findings.

A. Softening alone can improve the attacker's student

Softened teacher outputs are the classic recipe for knowledge distillation [2]. However, whether softening assists the adversary is governed by how much dark knowledge the teacher has to reveal.

On CIFAR-100, with no poison at all ($\lambda = 0$), serving at $T = 4$ (median served confidence 0.42) *improves* the distilled student by $+4.7$ pp over the undefended baseline ($T = 2$: $+0.9$ pp). On CIFAR-10 the benefit is absent: pushing to higher temperatures only degrades the student, monotonically, from -1.0 pp at $T = 4$ to -6.5 pp at $T = 12$. At $T = 10$, CIFAR-10 reproduces CIFAR-100's $T = 4$ operating point (served median confidence 0.42 on both), yet CIFAR-100 gains $+4.7$ pp where CIFAR-10 loses -5.2 pp. Tiny-ImageNet interpolates between the two exactly as this account predicts: its non-target mass is intermediate (Table 1), and so is the benefit gained by temperature scaling, that is, $+1.6$ pp at $T = 4$, which lies between CIFAR-10's absent benefit and CIFAR-100's $+4.7$ pp. At matched served confidence the sign of the effect is therefore set by how much usable non-target structure the teacher has to reveal, not by the magnitude of the temperature parameter or by the class count. Wherever dark knowledge exists, revealing it assists the adversary; where it does not, softening only dilutes the supervision the attacker already had. Note that constant-T scaling preserves the argmax, so this assistance is not even purchased with teacher accuracy: its only cost is served fidelity, which is the axis on which Sec. VI-D prices every softening variant.

B. Softening activates ADS inside the inert window

Composed with the poison, softening behaves exactly as (7) predicts: it shifts the entire transition curve into the formerly inert window. At $\lambda = 0.2$—inert under plain serving—the poison's marginal effect on the student (measured relative to the same serving at $\lambda = 0$, which nets out the shift of Sec. VI-A) grows from -1.1 pp (unsoftened) to -3.0 pp ($T = 2$) to -17.8 pp ($T = 4$) on CIFAR-100, and from -0.4 to -5.8 to -27.7 pp on CIFAR-10 (Fig. 7a): the same small λ becomes $17\times$–$69\times$ more potent. This provides direct causal evidence that the availability of non-target mass, not the penalty strength, is what limits the poison at low λ. CIFAR-10 sharpens the point: the same $T = 4$ serving that conferred no benefit on a clean student (Sec. VI-A) multiplies the poison's potency $69\times$—what softening manufactures is probability mass for the poison to reweight, whether or not that mass encodes anything a student could exploit. The teacher-side cost agrees with the theory quantitatively: the pre-registered prediction $F(\tau\lambda)$ placed the $T = 4$, $\lambda = 0.2$ teacher cost at 26.9 pp (CIFAR-100) and 30.1 pp (CIFAR-10); the trained runs measured 25.1 and 29.1 pp.

C. The same rescaling collapses the teacher

The teacher-side cost just quoted is the other half of the same mechanism: softening activates the poison by shrinking the

very margins the poison then crosses, so enablement and collapse are one phenomenon, not two. On the student-versus-teacher-cost plane (Fig. 7b), every constant-temperature variant ($T = 1$, 2, 4, on both datasets) collapses onto a single trade-off curve. Temperature never improves the exchange rate rather it only moves the operating point further along it. A worked example makes the severity concrete. A typical softened CIFAR-100 row has served margin 0.38 against poison contrast 16.8, so its prediction flips at $\lambda^* \approx 0.022$; measured end-to-end, $T = 4$ serving loses 25.1/29.1 pp of teacher accuracy (CIFAR-100/CIFAR-10) at $\lambda = 0.2$, a strength that is inert under plain serving, and 45.0/56.3 pp at $\lambda = 0.35$ (Fig. 7). Softening manufactures precisely the thin margins on which a margin-blind poison is most destructive because the two mechanisms compete for the same resource, since the softener consumes the margin the poison needs in order to act safely.

D. At matched served fidelity, softening is dominated

Constant temperature is the bluntest possible softener, so its failure does not by itself close the question. We therefore construct the strongest softening variant we could: a *gated* operator, inspired by [59], that lowers the confidence of over-confident inputs only, and does so by lowering the top logit alone, leaving all non-target structure intact for the dark-knowledge channel (Appendix XI-D). It is composed with the margin budget of Sec. VII, so that its teacher cost is identically zero. Section VIII-D evaluates this combination across five gates and the full λ grid against poison-only ADS-C on the axis a defender actually spends, served fidelity: the softener pays a large fidelity cost *up front*, before any poison flows, and at every matched fidelity level in the operating regime poison-only ADS-C damages the student more (Fig. 11). We therefore include no softening in the proposed defense: softening serves in this paper as the probe that establishes dark knowledge as the resource the poison consumes, and as a defense component it is dominated.

VII. ADS-C

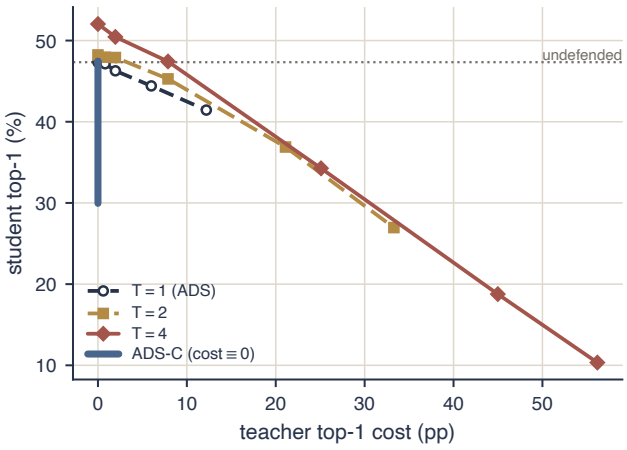
The diagnosis of Sec. V dictates the design requirement: retain the poison, but enforce margin-awareness at composition, per input. ADS-C implements this requirement in closed form.

A. The margin budget

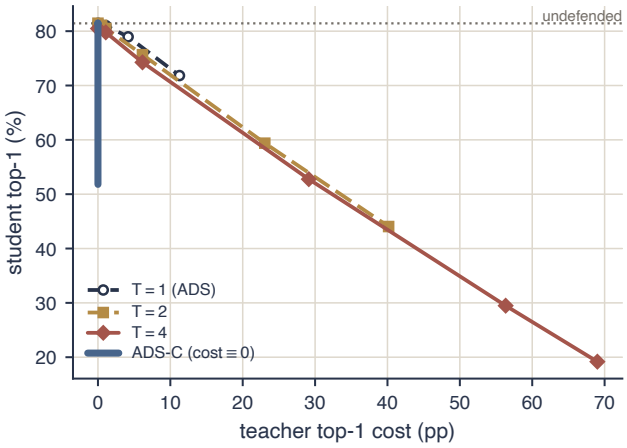
Fix an input $\mathbf{x}$ with teacher logits $\mathbf{z}$, top class $t = \arg\max_k z_k$, and poison $\widehat{\Delta}$. For every rival $j \neq t$, let $\mathrm{gap}_j = z_t - z_j$ denote the served margin to that rival and recall from (4) the rate $c_j = \widehat{\Delta}_j - \widehat{\Delta}_t$ at which the poison closes it. Given the requested strength λ and a margin floor $m > 0$, the per-input budget is

$$\lambda(\mathbf{x}) \;=\; \mathrm{clip}\Big(\min_{j:\, c_j > 0} \frac{\mathrm{gap}_j - m}{c_j} \,,\; 0,\; \lambda \Big), \qquad (8)$$

with the convention that a minimum over an empty set is $+\infty$, and the API serves $\hat{\mathbf{p}} = \sigma\big(\log \mathbf{p}_{\mathcal{T}} + \lambda(\mathbf{x})\,\widehat{\Delta}\big)$



(a) CIFAR-100



(b) CIFAR-10

FIGURE 7. Constant-temperature serving ± ADS on the student-versus-teacher-cost trade-off plane (E-CT; 3 seeds). Every constant-temperature arm (T=1, 2, 4) collapses onto a *single* unfavorable curve: softening activates the poison—at $\lambda = 0.2$ the T=4 poison is $17\times/69\times$ more potent than unsoftened, in quantitative agreement with the pre-registered $F(\tau\lambda)$ prediction—but only by shrinking the margins the poison then crosses, so the extra student damage is bought at proportional teacher cost. ADS-C (vertical line, teacher cost $\equiv 0$ by construction) reaches student damage no temperature configuration attains without spending the teacher accuracy the softener pays up front.

(Algorithm 1). The budget is the flip threshold (5) computed with m log-units of every margin held in reserve, clipped to the feasible range, and its behavior at the extremes is the intended one: rivals the poison pushes down ($c_j \leq 0$) never bind, and an input none of whose rivals gain passes the full requested strength. Conversely, an input whose margin is already at or below the floor receives no poison rather than a flip. Figure 8 traces one query through the full pipeline.

B. The guarantee

Proposition 1 (Zero-cost argmax preservation):

Algorithm 1 ADS-C

Require: query $\mathbf{x}$; teacher $\mathcal{T}$; frozen proxy clones $\mathcal{P}_+, \mathcal{P}_-$
Require: step ε; requested strength λ; margin floor m
Ensure: served probability vector $\hat{\mathbf{p}}$
$\mathbf{z} \leftarrow \mathcal{T}(\mathbf{x})$; $\quad t \leftarrow \arg\max_k z_k$
$\widehat{\Delta} \leftarrow \big[\log\sigma(\mathcal{P}_+(\mathbf{x})) - \log\sigma(\mathcal{P}_-(\mathbf{x}))\big]/(2\varepsilon)$
for $j \neq t$ **do**
  $\mathrm{gap}_j \leftarrow z_t - z_j$; $\quad c_j \leftarrow \widehat{\Delta}_j - \widehat{\Delta}_t$
end for
$\lambda(\mathbf{x}) \leftarrow \mathrm{clip}\big(\min_{j:c_j>0}(\mathrm{gap}_j - m)/c_j,\ 0,\ \lambda\big)$
return $\hat{\mathbf{p}} \leftarrow \sigma\big(\log\sigma(\mathbf{z}) + \lambda(\mathbf{x})\,\widehat{\Delta}\big)$

For every input $\mathbf{x}$ with unique teacher argmax t, the served vector of Algorithm 1 satisfies, for all $j \neq t$,

$$\log\hat{p}_t - \log\hat{p}_j \;\geq\; \min\big(m,\ \mathrm{gap}_j\big) \;>\; 0.$$

Consequently $\arg\max_k \hat{p}_k = \arg\max_k p_{\mathcal{T},k}$ for every query, and the defended teacher's top-1 accuracy equals the undefended teacher's top-1 accuracy identically—as a deterministic, per-query property, not in expectation. Moreover, $\lambda(\mathbf{x})$ is maximal: it is the largest value in $[0, \lambda]$ for which the stated margin bound holds.

Proof:
Softmax composition preserves log-score differences, so $\log\hat{p}_t - \log\hat{p}_j = \mathrm{gap}_j - \lambda(\mathbf{x})\,c_j$. If $c_j \leq 0$ this is $\geq \mathrm{gap}_j$. If $c_j > 0$, then by (8) $\lambda(\mathbf{x}) \leq (\mathrm{gap}_j - m)/c_j$ whenever that bound is nonnegative, giving $\mathrm{gap}_j - \lambda(\mathbf{x})c_j \geq m$; when the bound is negative (i.e., $\mathrm{gap}_j < m$), the clip yields $\lambda(\mathbf{x}) = 0$ and the difference equals gap_j. In all cases the difference is $\geq \min(m, \mathrm{gap}_j) > 0$, so t remains the served argmax. Maximality: constraints with $c_j \leq 0$ hold for every $\lambda' \geq 0$; for $c_j > 0$, the constraint $\mathrm{gap}_j - \lambda' c_j \geq \min(m, \mathrm{gap}_j)$ reads $\lambda' \leq (\mathrm{gap}_j - m)/c_j$ when $\mathrm{gap}_j \geq m$ and $\lambda' \leq 0$ when $\mathrm{gap}_j < m$. The feasible subset of $[0, \lambda]$ is therefore the interval $\big[0,\ \mathrm{clip}\big(\min_{j:c_j>0}(\mathrm{gap}_j - m)/c_j,\ 0,\ \lambda\big)\big]$, whose right endpoint is exactly (8). ∎

Remark 1:
The guarantee is unconditional: it requires no retraining, no access to training data, and no distributional assumption, and it holds for every input, including out-of-distribution queries. Empirically we confirm it to machine precision: across all 360 evaluated (dataset, variant, λ, seed) configurations, the largest observed teacher top-1 deviation is 1.4×10^{-6} pp—orders of magnitude below the 10^{-2} pp contribution of a single test image, i.e., floating-point noise in the accuracy average rather than any flipped prediction (Sec. VIII).

C. Design discussion

The floor m. The floor is the service level the defender guarantees on served margins: every served prediction retains at least $\min(m, \text{original margin})$ log-units of separation. We use $m = 0.05$ throughout; the guarantee itself is m-independent (teacher drop is zero for every m), and Appendix XI-C shows the poison passed through the budget is insensitive over $m \in [0.01, 0.2]$ (served poison-KL varies by $\approx$10% across the range).

Overconfidence cuts both ways. The same heterogeneity that made the global penalty fatal makes the budget cheap: most inputs sit far from the floor, so most of the requested strength survives. At $\lambda = 0.5$ the mean effective strength is 0.44 (CIFAR-100) and 0.47 (CIFAR-10), corresponding to 88–95% of the request, while the thin-margin tail, which a global λ would flip first, is individually protected. The budget spends the poison precisely where the confidence structure of Fig. 2 has room for it.

Deployment properties. ADS-C is inference-time only. Per query, beyond the teacher's forward pass, it costs two forward passes through the frozen proxy clones, each substantially smaller than the teacher, and $O(K)$ arithmetic for (8). The served output remains a valid probability vector with unchanged argmax. Like ADS, and unlike static perturbation defenses [25], [26], the perturbation is input-dependent and computed from state the attacker cannot observe (the proxy and its gradient g), so there is no fixed transformation to invert. Additionally, unlike ADS, its utility cost is identically zero.

VIII. EXPERIMENTS

A. Setup

Benchmarks. CIFAR-100 (teacher ResNet-110 $\rightarrow$ student ResNet-20), CIFAR-10 (ResNet-56 $\rightarrow$ ResNet-20), and Tiny-ImageNet ($K = 200$, 64×64; teacher ResNet-50, ImageNet-pretrained and fine-tuned with a small-input stem $\rightarrow$ student ResNet-18). The proxy always matches the student architecture; Appendix XI-E reports training recipes. Undefended anchors: CIFAR-100 teacher 71.23%, distilled student 47.34 ± 0.37%; CIFAR-10 teacher 92.70%, student 81.44 ± 0.43%; Tiny-ImageNet teacher 80.07%, student 47.84 ± 0.59%.

Attacker. The fixed distillation attacker of Sec. III: it queries the API once per image with the full unlabeled training split (50 000 queries on the CIFAR datasets, 100 000 on Tiny-ImageNet), unaugmented and at native resolution, then trains the student for 50 epochs on the returned (image, served-vector) pairs with the pure soft-label loss (1) at temperature 1 (full optimizer settings in Appendix XI-E). Every configuration is run over three seeds ($\{92, 786, 1337\}$); we report mean $\pm$ std.

Defense grid. $\lambda \in \{0, 0.1, 0.2, 0.35, 0.5, 0.75, 1, 1.5, 2, 3\}$; margin floor $m = 0.05$; $\varepsilon = 3 \times 10^{-3}$ (CIFAR) and 10^{-2} (Tiny-ImageNet; Appendix XI-B). *ADS* denotes the unmodified transfer (3); *ADS-C* denotes Algorithm 1; *ADS-C + softener* denotes the ablation of Sec. VIII-D; the guarantee relaxations q and $\lambda_{\min}$ (Sec. IX) default to $q = 1$, $\lambda_{\min} = 0$.

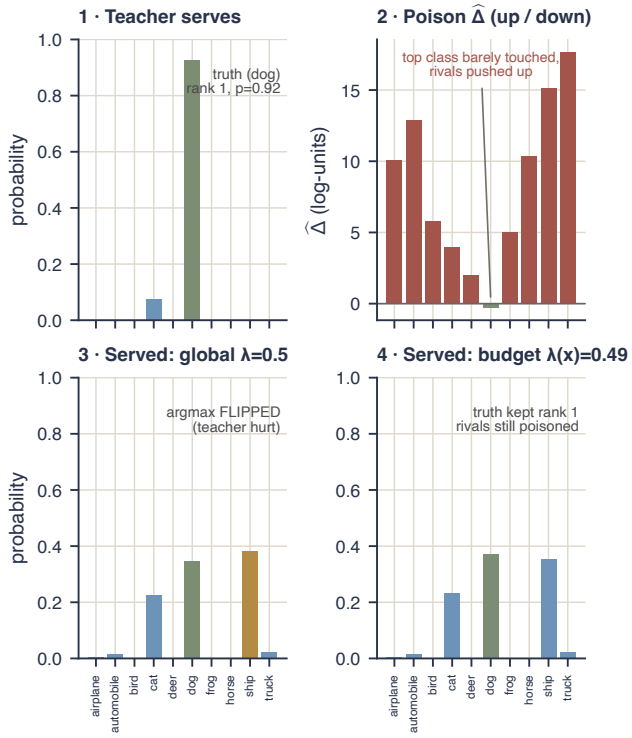


FIGURE 8. **One CIFAR-10 query through the pipeline (panels 1–4). The teacher serves a confident vector (panel 1; truth *dog*, $p_{\max}=0.92$); the poison $\widehat{\Delta}$ leaves the top class essentially untouched ($\widehat{\Delta}_t=-0.3$) while pushing every rival up, hardest on deep-ranked classes (panel 2); a global $\lambda = 0.5$ flips the served argmax to the hard-pushed *ship* (panel 3, teacher hurt), while the margin budget (8) spends $\lambda(\mathbf{x}) = 0.49$—just short of the flip—leaving the top-1 intact and the dark knowledge corrupted (panel 4). Pushes surface in the served vector only where the teacher leaves probability mass (*ship*, *cat*); equally large pushes on near-massless classes remain invisible at this strength—the starvation of Sec. V-A in miniature.**

TABLE 2. **Representative operating points (mean±std over 3 seeds). Teacher drop for ADS-C is exactly 0 by Proposition 1 (observed $\leq 1.4 \times 10^{-6}$ pp).**

	Defense	λ	$\mathcal{S}$ top-1 (%)	$\mathcal{T}$ drop (pp)
CIFAR-100	None	—	47.34 ± 0.37	0.00
	ADS	0.5	41.44 ± 0.09	12.19
	ADS	3	4.21 ± 0.05	64.01
	ADS-C	1	37.16 ± 0.27	**0.00**
	ADS-C	3	$\mathbf{29.95 \pm 0.23}$	**0.00**
	Hard-label	—	47.01 ± 0.64	0.00
CIFAR-10	None	—	81.44 ± 0.43	0.00
	ADS	0.5	71.83 ± 0.27	11.26
	ADS	3	12.94 ± 0.05	77.61
	ADS-C	1	64.74 ± 0.10	**0.00**
	ADS-C	3	$\mathbf{51.80 \pm 0.38}$	**0.00**
	Hard-label	—	81.55 ± 0.47	0.00
Tiny-ImageNet	None	—	47.84 ± 0.59	0.00
	ADS	1	43.29 ± 0.20	8.97
	ADS	3	19.09 ± 0.15	46.65
	ADS-C	1	45.19 ± 0.41	**0.00**
	ADS-C	3	$\mathbf{34.51 \pm 0.28}$	**0.00**

B. Main result: the trade-off rate before and after budgeting

Figure 9 plots student accuracy against teacher accuracy for both methods on all three benchmarks; Table 2 reports representative operating points.

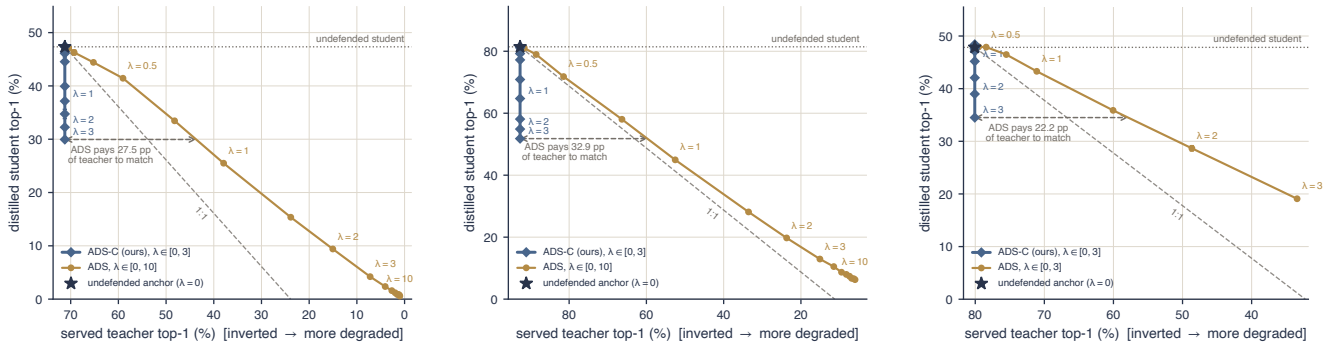
Unmodified ADS. The direct transfer degrades the student only by degrading the teacher comparably or faster. On CIFAR-100 the sweep passes through $(-12.2\,\text{pp}$ teacher, $-5.9\,\text{pp}$ student) at $\lambda = 0.5$ and ends at $(-64.0\,\text{pp}, -43.1\,\text{pp})$ at $\lambda = 3$. The teacher suffers more than the student. Tiny-ImageNet exhibits the same shape shifted by its wider inert window (Sec. V-D). Its exchange is still near-inert at $\lambda = 0.5$ ($-1.7\,\text{pp}$ teacher, $0.0\,\text{pp}$ student), first becomes active at $\lambda = 1$ ($-9.0\,\text{pp}$ teacher for $-4.6\,\text{pp}$ student, rate 0.51), and reaches $(-46.7\,\text{pp}, -28.8\,\text{pp})$ at $\lambda = 3$.

ADS-C. With the budget active, teacher top-1 drop is $0.00\,\text{pp}$ at every cell, while the student falls monotonically with λ. It falls to $29.95 \pm 0.23\%$ on CIFAR-100 ($-17.4\,\text{pp}$, 37% relative), $51.80 \pm 0.38\%$ on CIFAR-10 ($-29.6\,\text{pp}$, 36% relative), and $34.51 \pm 0.28\%$ on Tiny-ImageNet ($-13.3\,\text{pp}$, 28% relative) at $\lambda = 3$.

Matched-damage comparison. For unmodified ADS to inflict the student damage that ADS-C delivers at zero cost, it must spend $27.5\,\text{pp}$ (CIFAR-100), $32.9\,\text{pp}$ (CIFAR-10), and $22.2\,\text{pp}$ (Tiny-ImageNet) of teacher accuracy. The margin budget converts an unfavorable trade into a free one.

C. Decomposition of the unbudgeted poison's damage

The budget also functions as an analytical instrument. It separates the unbudgeted poison's student damage into an argmax-safe component, which survives the budget, and flip collateral, which the budget removes (Fig. 10). At matched $\lambda = 0.5$ on CIFAR-100, the unbudgeted poison's $5.9\,\text{pp}$ student drop decomposes into $2.8\,\text{pp}$ of argmax-safe corruption (retained at zero teacher cost) and $3.1\,\text{pp}$ of flip collateral (purchased at $12.2\,\text{pp}$ of teacher accuracy); on CIFAR-10, $5.0\,\text{pp}$ of $9.6\,\text{pp}$ survives (52%). At smaller strengths the argmax-safe share is larger still (87% at $\lambda = 0.2$ on CIFAR-100). Two conclusions follow. First, roughly half of the student damage that unmodified ADS purchased with teacher accuracy was genuine dark-knowledge corruption that required no payment at all. Second, because the cost ceiling is removed, ADS-C can increase λ far past the point where unmodified ADS becomes unusable, more than recovering the remainder: its zero-cost $-17.4\,\text{pp}$ (CIFAR-100) / $-29.6\,\text{pp}$ (CIFAR-10) at $\lambda = 3$ exceeds the unbudgeted poison's total damage at any teacher cost below 27–33 pp. This decomposition is, to our knowledge, the first quantitative account of *how* ADS corrupts dark knowledge



(a) CIFAR-100 (b) CIFAR-10 (c) Tiny-ImageNet

FIGURE 9. The main result: distilled-student top-1 versus served-teacher top-1 (horizontal axis inverted—the teacher degrades left to right; 3 seeds; the swept λ is annotated along each curve, with ADS extended to $\lambda = 10$ on the CIFAR panels; dashed gray: the 1:1 break-even through the undefended anchor $\star$). Unmodified ADS traces a diagonal: it degrades the student only by degrading the teacher comparably or faster. ADS-C is the vertical segment at zero teacher cost (Proposition 1). The horizontal arrow marks the teacher accuracy unmodified ADS must spend to match ADS-C's free student damage at $\lambda = 3$.

FIGURE 10. Decomposition of unmodified ADS damage at matched λ. The unbudgeted student drop splits into an argmax-safe part (survives the budget, at zero cost) and flip collateral (purchased at 11–12 pp of teacher accuracy).

in classification: the durable damage is corruption of non-target structure, not label noise. This conclusion is further corroborated by Sec. IX-B from an independent direction.

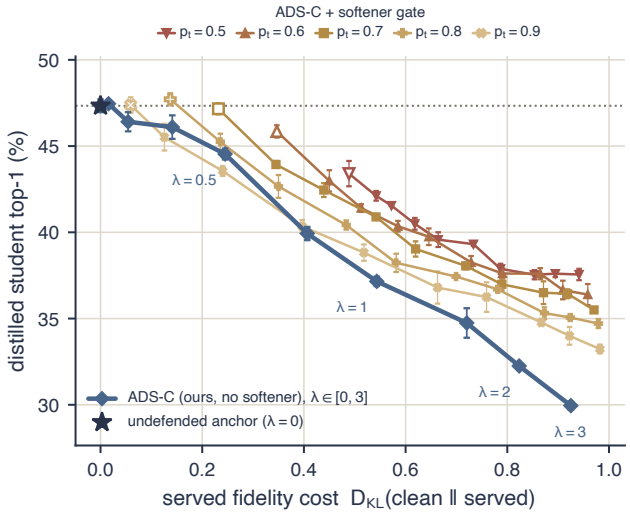
D. Ablation: softening front-ends at matched served fidelity

Finally we place the strongest softening variant, namely the gated top-logit reduction of Sec. VI-D, composed with the ADS-C against ADS-C with no temperature scaling. We sweep the softener's gate over five settings and λ over the full grid (360 runs total across both datasets; teacher drop $\equiv 0$ everywhere, as guaranteed).

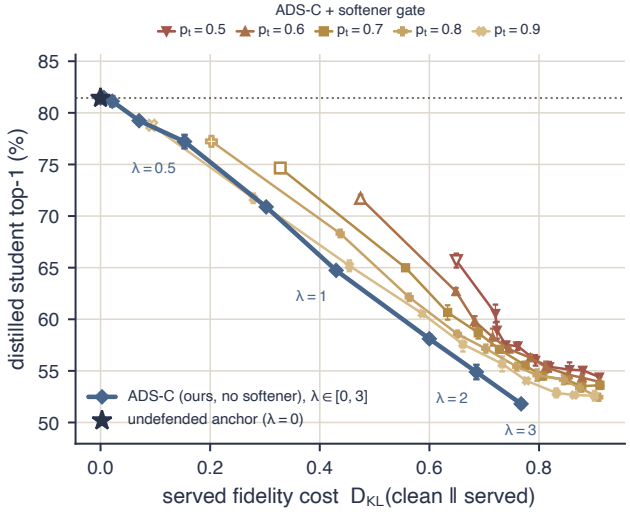
At matched λ the comparison is misleading as the softener is stronger at low λ (by up to 6.3 pp on CIFAR-100 and 22.2 pp on CIFAR-10) and weaker at high λ. But λ is free. The axes on which a defender actually pays are teacher drop (identically zero for both variants) and served fidelity. Recomputed on the fidelity axis (Fig. 11), the comparison inverts as the softener spends a large fidelity budget up front ($D_{KL} = 0.49$ at $\lambda = 0$ for its most aggressive gate) before any poison flows, while ADS-C spends fidelity only as λ grows. At matched served-KL, ADS-C without any softening damages the student more everywhere in the operating regime. Its advantage widens with fidelity spend, from $+1.0$ pp at KL 0.5 to $+3.8$ pp at KL 0.9 on CIFAR-100, and $+1.0$ to $+2.2$ pp across KL 0.4–0.7 on CIFAR-10. ADS-C's deepest zero-cost damage ($29.95/51.80$) is unmatched by any softener configuration on either dataset. The softener retains a ≤ 1 pp advantage only in the low-damage corner (KL ≤ 0.3) that is undesirable for defense. The matched-λ crossover is thus an artifact of comparing on an axis that costs nothing; at matched served fidelity, softening is dominated. This ablation also answers the natural question of why ADS-C includes no confidence-scaling component: the strongest such component we constructed is dominated by not including one.

E. Controls

The hard-label attacker. We train the attacker of Sec. III on the served argmax alone (cross-entropy; implemented as KL against the one-hot served label, which is identical). Under ADS-C the served argmax equals the clean argmax on every query. We verify zero deviations over both $50\,000$-query corpora at $\lambda \in \{2, 3\}$. Therefore this attacker receives identical supervision from the defended and undefended APIs. Two observations follow. First, these overconfident teachers serve vectors so nearly one-hot that, without softening, dark knowledge confers almost no advantage to begin with, consistent with Table 1, and complementary to Sec. VI-A, where the advantage appears ($+4.7$ pp) as soon as softening reveals the non-target structure. Second, and decisive for the



(a) CIFAR-100



(b) CIFAR-10

FIGURE 11. The served-fidelity trade-off: distilled-student accuracy versus served $D_{KL}(\mathbf{p}_\mathcal{T}\|\hat{\mathbf{p}})$; teacher drop $\equiv 0$ for every point (mean±std whiskers over 3 seeds; $\star$ marks the undefended λ=0 anchor). Each softener gate starts already displaced right (open markers): fidelity spent up front, at λ=0, before any poison flows—up to $D_{KL} = 0.49$ (CIFAR-100) and 0.65 (CIFAR-10) for the most aggressive gate p_t=0.5. Poison-only ADS-C is the lower envelope throughout the operating regime, and at the deepest fidelity it spends (λ=3) the best softener configuration at matched served-KL leaves the student 4.0 pp / 2.6 pp higher (arrows): softening never recovers its up-front spend.

defense: ADS-C drives the soft-label student accuracy down by 17.1 pp (CIFAR-100, $\lambda = 3$) and 29.7 pp (CIFAR-10). The defended soft output is strictly worse distillation signal than the argmax it accompanies. The defense does not merely remove the incentive to distill from served probabilities, it reverses it, while the argmax channel remains exactly as informative as any accurate API must be. Sec. IX addresses the attacker who exploits knowledge of the guarantee itself.

F. Fidelity of the stolen copy

Top-1 accuracy understates the damage to an attacker whose goal is a faithful replica rather than a merely accurate one (e.g., to mount downstream evasion or membership-inference attacks [13], [15]). Agreement between student and teacher predictions falls in step with student accuracy (CIFAR-100: 0.49 undefended $\rightarrow 0.31$ at $\lambda = 3$; CIFAR-10: $0.83 \rightarrow 0.52$): the extracted model is not only worse, it is a worse copy.

IX. TRADING THE GUARANTEE FOR ATTACKER UNCERTAINTY

Proposition 1 is deliberately absolute, and absoluteness is itself information: an attacker who knows the defense knows that every served top-1 is honest, and that inputs whose margin lies at or below the threshold are served identically to the clean teacher. Those provably clean rows are, moreover, exactly the thin-margin inputs where near-boundary class structure, corresponding to the richest dark knowledge, resides. A defender may therefore wish to spend a small, controlled amount of the guarantee to remove these certainties. We provide two such relaxations, each with closed-form-predictable utility cost and each disabled by default; ADS-C with $q = 1$ and $\lambda_{\min} = 0$ remains the proposed method. All results below are trained-student measurements (3 seeds, both CIFAR datasets).

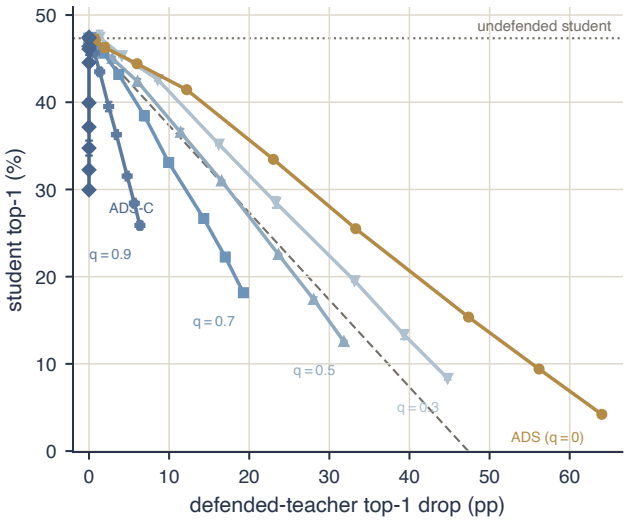
A. Stochastic enforcement of the budget

The first relaxation enforces the budget (8) on each input with probability q; with probability $1 - q$ the input receives the full requested λ (raw poison). The endpoints recover the two extremes—$q = 1$ is ADS-C, $q = 0$ is unmodified ADS—and the expected teacher cost follows in closed form from Sec. V-D: $(1 - q)$ times the unmodified-ADS teacher cost at λ predicted by (6), so the defender can price the relaxation from cached outputs before deploying it. The enforcement mask is drawn once per input from a fixed seed, not per query (Sec. IX-C).

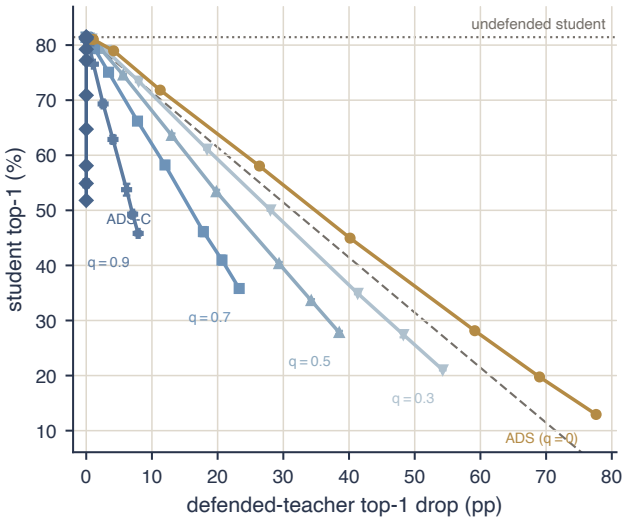
Figure 12 shows the resulting family of trade-off curves, which interpolates between the ADS-C vertical and the ADS diagonal. The trade-off is ordered *exactly* by q on both datasets: at every level of teacher cost, the best achievable student damage is obtained by the largest q whose curve reaches that cost, so no configuration with smaller q is ever preferable at equal spend. The exchange below the zero-cost floor is far better than unmodified ADS offers: relaxing to $q = 0.9$ at $\lambda = 3$ yields a further -4.1 pp of student damage (CIFAR-100, to 25.9%) for 6.4 pp of teacher cost, where unmodified ADS would require 32.8 pp to match—approximately a $5\times$ better exchange; on CIFAR-10, -6.0 pp for 7.9 pp against 39.3 pp ($5.0\times$). Extending the sweep to $\lambda = 10$ on CIFAR-100 strengthens this to a uniform statement: the relaxed configurations saturate ($q = 0.9$ reaches student accuracy 22.7% at 6.9 pp, more than one student-pp per teacher-pp below the floor), and unmodified ADS becomes dominated at *every* point of the trade-off—

its deepest operating point (student 4.2% at 64.0 pp) is surpassed by $q = 0.3$ at $\lambda = 10$ (student 3.5% at 49.1 pp).

The hard-label measurements of Sec. VIII-E add an important qualification (Fig. 13). Once $q < 1$, the two attacker channels separate, and at every q the *hard-label* student is the stronger one (38.5% versus 25.9% at $q = 0.9$ on CIFAR-100; 73.1 versus 45.8% on CIFAR-10): a defense-aware attacker responds to the relaxation by distilling from the served argmax, whose label noise—the flipped rows—arrives at the closed-form rate $(1-q)F(\lambda)$. Priced against this best response, the exchange below the floor is 1.35, 1.02, 0.89, and 0.81 student-pp per teacher-pp at $q = 0.9$, 0.7, 0.5, and 0.3 on CIFAR-100 (1.07, 1.02, 0.99, and 0.97 on CIFAR-10): approximately break-even near the guarantee, drifting below it at deep relaxation, and in every case far from the $5\times$ available against an attacker committed to soft labels. Three observations delimit the relaxation's value. First, the best-responding attacker still degrades below the floor at every q—the deliberate flips are systematic, input-consistent noise that the student learns rather than averages away—so no channel recovers the undefended leakage. Second, the exchange rates place the defensible operating points at shallow relaxation ($q \gtrsim 0.7$), where the trade stays at or above break-even against either channel; deeper relaxation is justified only against an attacker known to consume soft labels. Third, the guarantee point $q = 1$ remains uniquely favorable: it is the only setting at which the defender pays nothing while the attacker's best response yields exactly the floor.



(a) CIFAR-100



(b) CIFAR-10

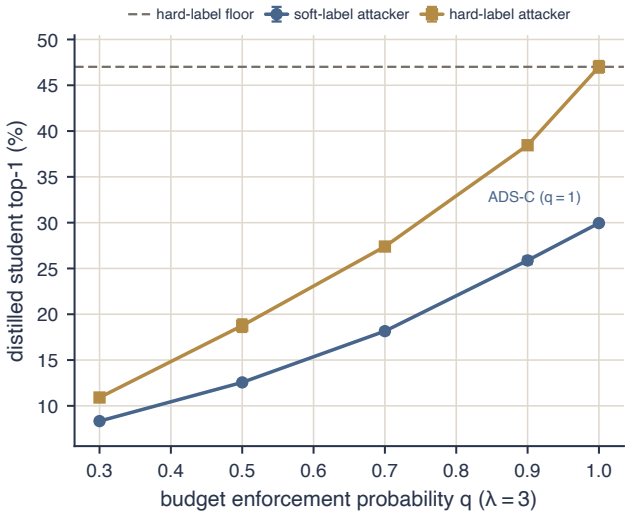
FIGURE 12. Stochastic enforcement of the budget (margin cap enforced with probability q; 3 seeds; dashed gray: 1:1 break-even). The family of trade-off curves interpolates between ADS-C ($q = 1$, vertical) and ADS ($q = 0$, diagonal) and is ordered exactly by q on both datasets. Below ADS-C's zero-cost floor, $q = 0.9$ obtains additional student damage at approximately $5\times$ better exchange rate than unmodified ADS against the soft-label attacker; Fig. 13 prices the same relaxation against the attacker's best response.

B. Defensive flipping with a minimum λ

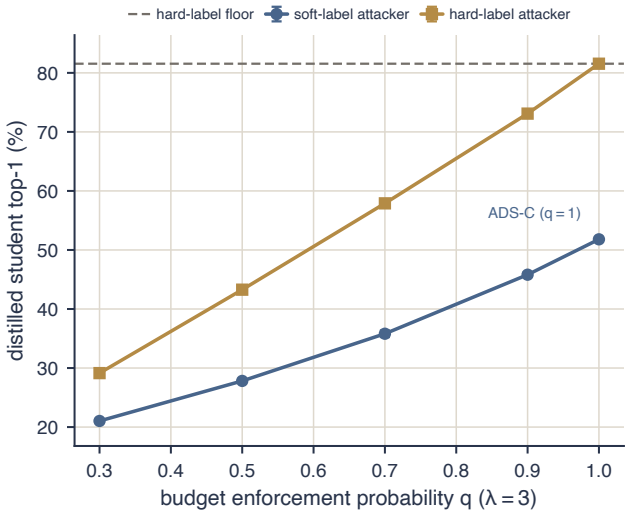
The second relaxation addresses the clean-row exposure directly: serve every input with strength at least $\lambda_{\min}$, i.e., $\lambda_{\text{eff}}(\mathbf{x}) = \max(\lambda_{\min}, \lambda(\mathbf{x}))$. Inputs whose flip threshold lies below $\lambda_{\min}$ are deliberately flipped, which we refer to as *defensive flipping*, concentrated by construction on the thinnest-margin rows, where flips are least costly. The gross cost is again available in closed form from the transition curve, $F(\lambda_{\min})$, and the *net* cost is smaller, for a geometric reason: thin-margin rows are disproportionately already misclassified, and on such rows the true class is often the runner-up (68% of flipped misclassified rows on CIFAR-10, 33% on CIFAR-100), so a fraction of defensive flips land on the true label and partially offset the gross cost (net cost $\approx \frac{1}{2}$ gross on CIFAR-100; the offset weakens with K). This is a consequence of margin geometry rather than of any corrective aim in the poison, and we report it as such.

Measured end-to-end, the minimum-strength floor serves a single purpose, at near-zero cost: at $\lambda_{\min} = 0.05$, $\lambda = 3$, *no served row is clean* (the clean-row fraction falls from 0.4% to ≈0) at a teacher cost of 0.30 pp (CIFAR-100) / 0.16 pp (CIFAR-10), with student accuracy statistically indistinguishable from the ADS-C floor (30.1 vs. 29.95; 51.4 vs. 51.80). It is a hardening mechanism—it removes the attacker's certainty about which rows are unperturbed—not a means of increasing damage. Nor does it open a side channel: the hard-label attacker under $\lambda_{\min} \in \{0.05, 0.2\}$ remains within seed noise of its leakage floor (46.7/47.3 vs. 47.0% on CIFAR-100; 81.3/81.5 vs. 81.6% on CIFAR-10), as the $\leq 0.2\%$ label-noise rate predicts.

A stronger variant is conceivable: sacrificing the thin-margin rows *entirely* (serving them the full unbudgeted λ) is the margin-ranked analogue of stochastic enforcement, and on served-distribution corruption (poison-KL) it appeared to dominate stochastic sacrifice by a factor of ∼2 in our

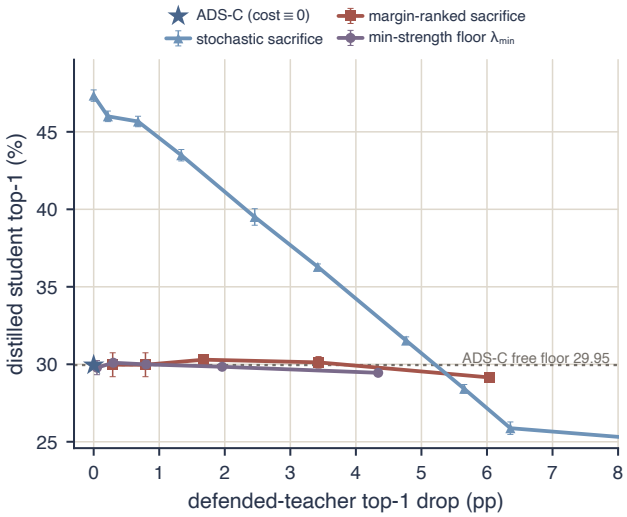


(a) CIFAR-100

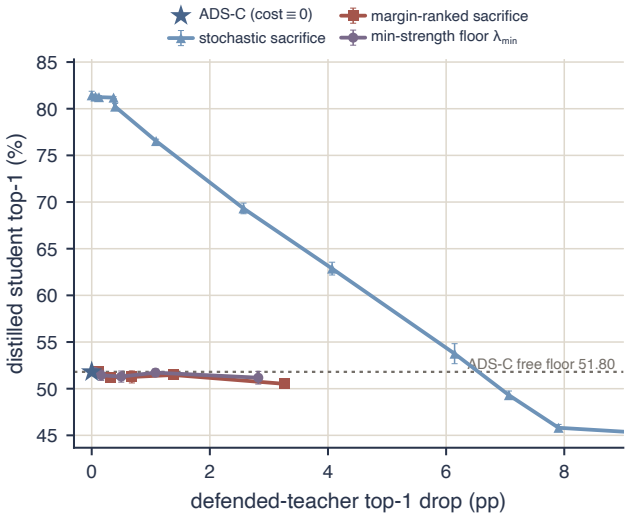


(b) CIFAR-10

FIGURE 13. **The two attacker channels under stochastic enforcement ($\lambda = 3$, 3 seeds): distilled-student accuracy of the soft-label attacker (served probabilities) and the hard-label attacker (served argmax) as a function of the enforcement probability q. At $q = 1$ the argmax channel is identical to the undefended API floor (dashed) while the soft channel lies $17.1\,\mathrm{pp}$ (CIFAR-100) and $29.7\,\mathrm{pp}$ (CIFAR-10) below it. At every $q < 1$ the attacker's best response flips to hard labels, whose deliberate label noise arrives at the closed-form rate $(1 - q)F(\lambda)$; both channels nevertheless remain below the floor throughout.**



(a) CIFAR-100



(b) CIFAR-10

FIGURE 14. **Defensive flipping versus stochastic sacrifice, in the low-cost region ($\star$: ADS-C at zero teacher cost; dotted rule: the ADS-C free floor; error bars ± 1 std over 3 seeds). Stochastic sacrifice (triangles; the q-relaxation of Sec. IX-A, drawn as its lower envelope over the q, λ sweep) descends below the floor; margin-ranked sacrifice (squares) is dominated—heavy poison on thin-margin rows inflates served KL divergence but not student damage; and the minimum-strength floor $\lambda_{\min}$ (circles), like margin-ranked shown at the deployment strength λ=3, sits at the ADS-C floor at $\leq 0.3\,\mathrm{pp}$ teacher cost while guaranteeing that no served row is clean.**

training-free pre-analysis. The trained students overturn this prediction (Fig. 14): margin-ranked sacrifice cannot push the student meaningfully below the ADS-C floor (best case 29.15% vs. floor 29.95% on CIFAR-100, at $6.0\,\mathrm{pp}$ teacher cost), whereas stochastic sacrifice at the same $\sim 6\,\mathrm{pp}$ reaches 25.9% and continues to 8.3% at higher spend. Concentrating heavy poison on near-boundary rows inflates the served KL divergence while teaching the attacker's student little it would not have learned regardless: those rows carry little *informative* dark knowledge relative to their contribution to the KL divergence. We report this negative result because it provides independent, converging evidence for the mechanism identified in Sec. VIII-C: what damages a distilling student is corruption of the informative inter-class structure on representative inputs, not maximization of distributional distortion where distortion is inexpensive.

C. Determinism and repeated queries

Any defense that randomizes *per query* is invertible by repetition: querying the same input many times and aggregating the responses recovers the clean signal. The mechanisms above are immune to this attack by construction. The base ADS-C perturbation is a deterministic function of the input (teacher logits and frozen proxy clones), so repeated queries return the identical perturbed vector; the enforcement mask of Sec. IX-A is drawn per *input* from a fixed seed, not per query, so a sacrificed input is consistently sacrificed; and $\lambda_{\min}$ is inherently per-input deterministic. Repeated queries therefore provide the attacker no additional information. The residual cost of determinism is a persistent error set under $\lambda_{\min}$ (the same flipped inputs are always flipped)—for a deployed API arguably preferable to nondeterministic answers—and identifying that set would require access to the clean teacher, which is the asset under protection. The utility statement weakens in form but remains strong: from "accuracy identical" to "accuracy drop at most $F(\lambda_{\min})$, known in closed form before deployment, and concentrated on inputs the teacher was already uncertain about."

X. DISCUSSION AND LIMITATIONS

What the guarantee does and does not cover. Proposition 1 protects the served argmax—the quantity consumed by classification API users. Applications that consume the full served distribution (e.g., downstream risk calibration) pay the served-fidelity cost, which we quantify explicitly: it is the horizontal axis of Fig. 11, and at the recommended operating point amounts to a served KL of 0.8–0.9 (CIFAR-100). In the dark-knowledge threat model this cost falls almost entirely on the attacker: the honest consumer of a classification API reads the label.

Attacker adaptivity. Our main results use the standard fixed distillation attacker. The hard-label attacker is reduced to the distillation floor every classification API shares. Repetition attacks are addressed by per-input determinism (Sec. IX-C). Attacks that estimate and subtract the poison would need to reconstruct a hidden gradient direction through a black-box interface; we view a formal analysis of this inversion problem as valuable future work.

Scale and generality. Our evidence is CIFAR-10, CIFAR-100, and Tiny-ImageNet. Nothing in ADS-C is architecture- or scale-specific, and the Tiny-ImageNet teacher already covers the deployment-realistic case of a strong pretrained model, whose overconfidence (Table 1) shows the phenomenon we exploit is a property of model quality rather than of small benchmarks. Validation on large-scale production teachers and modern architectures (e.g., vision transformers) remains future work; the transition-curve analysis of Sec. V-D provides a tool, computable from teacher and proxy outputs alone, to predict, before any training, how a new teacher will behave.

The trade-off-rate lens. We encourage future inference-time defenses to report (i) utility cost with the axes made explicit, and (ii) frontiers at matched served fidelity rather than matched internal knob settings; Sec. VIII-D shows the latter comparison can invert the former.

XI. CONCLUSION

We adapted Antidistillation Sampling to classification and showed that its behavior there is governed by teacher overconfidence: an inert window in which the defense measurably affects neither party, a transition curve that predicts the defended teacher's cost across the entire sweep, and a saturating regime in which the direct transfer trades below break-even. The perturbation itself is faithful and correctly aimed at dark knowledge; the failure is margin-blindness at composition, on teachers whose confidence is extremely heterogeneous. Temperature softening rescales the transition curve exactly as the theory predicts, which makes it a valuable probe and demonstrates that the availability of dark knowledge is what limits the perturbation at low strengths; but it is not a remedy: the same rescaling collapses the teacher, and every temperature configuration lies on the same unfavorable trade-off curve. ADS-C enforces the missing margin-awareness where both the margin and the perturbation are visible—at composition—with a closed-form per-input budget that provably preserves every served prediction. The result is, to our knowledge, the first antidistillation defense for classification with exactly zero utility cost: $17.4\,\mathrm{pp}$ (CIFAR-100), $29.6\,\mathrm{pp}$ (CIFAR-10), and $13.3\,\mathrm{pp}$ (Tiny-ImageNet) of distilled-student degradation at a teacher cost of $0.00\,\mathrm{pp}$. Measured against the complementary channel, the guarantee reverses the economics of extraction: an attacker restricted to the served argmax obtains exactly the distillation floor that every accurate classification API shares, while the defended soft output trains a student well below that floor. The served probabilities become strictly worse supervision than the labels they accompany. Around the guarantee we provide two relaxations with closed-form-predictable cost, stochastic enforcement and defensive flipping, that allow a defender to exchange a known amount of utility for attacker uncertainty, priced throughout against the attacker's best response across both supervision channels, together with a negative result showing that stochastic sacrifice outperforms margin-ranked sacrifice, further evidence that the durable damage resides in informative dark knowledge. A defense that costs nothing changes the deployment calculus as there is no longer a utility argument for serving undefended soft labels.

ACKNOWLEDGMENTS

The authors thank the MS Artificial Intelligence Program at the Rochester Institute of Technology for computational resources, and the Fulbright Program for scholarship support to the first author. M. J. Khojasteh acknowledges support from the Gleason Endowment and Provost's Learning Innovation Grant at RIT. The authors acknowledge Research Computing at the Rochester Institute of Technology for providing

computational resources and support that have contributed to the research results reported in this publication [60].

REFERENCES

[1] J. Schmidhuber, "Learning complex, extended sequences using the principle of history compression," *Neural computation*, vol. 4, no. 2, pp. 234–242, 1992.

[2] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[3] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[4] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[5] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357.

[6] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International journal of computer vision*, vol. 129, no. 6, pp. 1789–1819, 2021.

[7] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction {APIs}," in *25th USENIX security symposium (USENIX Security 16)*, 2016, pp. 601–618.

[8] K. Zhao, L. Li, K. Ding, N. Z. Gong, Y. Zhao, and Y. Dong, "A survey on model extraction attacks and defenses for large language models," in *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, 2025, pp. 6227–6236.

[9] CNBC, "DeepSeek trained AI model using distillation, now a disruptive force," CNBC, Feb. 2025, https://www.cnbc.com/2025/02/21/deepseek-trained-ai-model-using-distillation-now-a-disruptive-force.html.

[10] Anthropic, "Detecting and preventing distillation attacks," Anthropic News, February 2026, accessed: March 23, 2026. [Online]. Available: https://www.anthropic.com/news/detecting-and-preventing-distillation-attacks

[11] Y. Sun, T. Liu, P. Hu, Q. Liao, S. Fu, N. Yu, D. Guo, Y. Liu, and L. Liu, "Deep intellectual property protection: A survey," *arXiv preprint arXiv:2304.14613*, 2023.

[12] Y. Jiang, Y. Gao, C. Zhou, H. Hu, S. Chen, A. Fu, and W. Susilo, "Intellectual property protection for deep learning model and dataset intelligence," *Engineering Applications of Artificial Intelligence*, vol. 163, p. 113024, 2026.

[13] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE symposium on security and privacy (SP)*. IEEE, 2017, pp. 3–18.

[14] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes, "Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models," *arXiv preprint arXiv:1806.01246*, 2018.

[15] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, 2017, pp. 506–519.

[16] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli, "Evasion attacks against machine learning at test time," in *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 2013, pp. 387–402.

[17] D. Lowd and C. Meek, "Adversarial learning," in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, 2005, pp. 641–647.

[18] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart, "Privacy in pharmacogenetics: An {End-to-End} case study of personalized warfarin dosing," in *23rd USENIX security symposium (USENIX Security 14)*, 2014, pp. 17–32.

[19] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 2015, pp. 1322–1333.

[20] G. Ateniese, L. V. Mancini, A. Spognardi, A. Villani, D. Vitali, and G. Felici, "Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers," *International Journal of Security and Networks*, vol. 10, no. 3, pp. 137–150, 2015.

[21] C. J. Haug and J. M. Drazen, "Artificial intelligence and machine learning in clinical medicine, 2023," *New England Journal of Medicine*, vol. 388, no. 13, pp. 1201–1208, 2023.

[22] L. Hernandez Aros, L. X. Bustamante Molano, F. Gutierrez-Portela, J. J. Moreno Hernandez, and M. S. Rodríguez Barrero, "Financial fraud detection through the application of machine learning techniques: a literature review," *Humanities and Social Sciences Communications*, vol. 11, no. 1, pp. 1–22, 2024.

[23] S. Cao, M. Li, J. Hays, D. Ramanan, Y.-X. Wang, and L. Gui, "Learning lightweight object detectors via multi-teacher progressive distillation," in *International Conference on Machine Learning*. PMLR, 2023, pp. 3577–3598.

[24] Y. Savani, A. Trockman, Z. Feng, A. Schwarzschild, A. Robey, M. Finzi, and J. Z. Kolter, "Antidistillation sampling," *arXiv preprint arXiv:2504.13146*, 2025.

[25] T. Lee, B. Edwards, I. Molloy, and D. Su, "Defending against neural network model stealing attacks using deceptive perturbations," in *2019 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2019, pp. 43–49.

[26] Y. Chen, R. Guan, X. Gong, J. Dong, and M. Xue, "D-dae: Defense-penetrating model extraction attacks," in *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2023, pp. 382–399.

[27] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International conference on machine learning*. PMLR, 2017, pp. 1321–1330.

[28] C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 535–541.

[29] J. Ba and R. Caruana, "Do deep nets really need to be deep?" *Advances in neural information processing systems*, vol. 27, 2014.

[30] T. Furlanello, Z. Lipton, M. Tschannen, L. Itti, and A. Anandkumar, "Born again neural networks," in *International conference on machine learning*. PMLR, 2018, pp. 1607–1616.

[31] J. Tang, R. Shivanna, Z. Zhao, D. Lin, A. Singh, E. H. Chi, and S. Jain, "Understanding and improving knowledge distillation," *arXiv preprint arXiv:2002.03532*, 2020.

[32] T. Orekondy, B. Schiele, and M. Fritz, "Knockoff nets: Stealing functionality of black-box models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4954–4963.

[33] J.-B. Truong, P. Maini, R. J. Walls, and N. Papernot, "Data-free model extraction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4771–4780.

[34] S. Jahan and R. Sun, "Black-box behavioral distillation breaks safety alignment in medical llms," *arXiv preprint arXiv:2512.09403*, 2025.

[35] D. Oliynyk, R. Mayer, and A. Rauber, "I know what you trained last summer: A survey on stealing machine learning models and defences," *ACM Computing Surveys*, vol. 55, no. 14s, pp. 1–41, 2023.

[36] W. Jiang, H. Li, G. Xu, T. Zhang, and R. Lu, "A comprehensive defense framework against model extraction attacks," *IEEE Transactions on Dependable and Secure Computing*, vol. 21, no. 2, pp. 685–700, 2023.

[37] X. Xu, M. Li, C. Tao, T. Shen, R. Cheng, J. Li, C. Xu, D. Tao, and T. Zhou, "A survey on knowledge distillation of large language models," *arXiv preprint arXiv:2402.13116*, 2024.

[38] H. Ma, T. Chen, T.-K. Hu, C. You, X. Xie, and Z. Wang, "Undistillable: Making a nasty teacher that cannot teach students," *arXiv preprint arXiv:2105.07381*, 2021.

[39] E. Yilmaz and H. Y. Keles, "Adversarial sparse teacher: Defense against distillation-based model stealing attacks using adversarial examples," *IEEE Access*, 2025.

[40] C. Liang, J. Huang, Z. Zhang, and S. Zhang, "Defending against model extraction attacks with ood feature learning and decision boundary confusion," *Computers & Security*, vol. 136, p. 103563, 2024.

[41] X. Cheng, M. Zheng, S. Zhu, and Y. Dong, "Misleader: Defending against model extraction with ensembles of distilled models," *arXiv preprint arXiv:2506.02362*, 2025.

[42] M. Juuti, S. Szyller, S. Marchal, and N. Asokan, "Prada: protecting against dnn model stealing attacks," in *2019 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2019, pp. 512–527.

[43] M. Gurve, S. Behera, S. Ahlawat, and Y. Prasad, "Misguide: Defense against data-free deep learning model extraction," *arXiv preprint arXiv:2403.18580*, 2024.
[44] A. Chakraborty, S. F. Ahamed, S. Roy, S. Banerjee, K. Choi, A. Rahman, A. Hu, E. Bowen, and S. Shetty, "Radep: A resilient adaptive defense framework against model extraction attacks," in *2025 International Wireless Communications and Mobile Computing (IWCMC)*. IEEE, 2025, pp. 1241–1246.
[45] J. R. Douceur, "The Sybil attack," in *International Workshop on Peer-to-Peer Systems*. Springer, 2002, pp. 251–260.
[46] H. Yu, K. Yang, T. Zhang, Y.-Y. Tsai, T.-Y. Ho, and Y. Jin, "Cloudleak: Large-scale deep learning models stealing through adversarial examples," in *NDSS*, vol. 38, 2020, p. 102.
[47] Y. Uchida, Y. Nagai, S. Sakazawa, and S. Satoh, "Embedding watermarks into deep neural networks," in *Proceedings of the 2017 ACM on international conference on multimedia retrieval*, 2017, pp. 269–277.
[48] H. Jia, C. A. Choquette-Choo, V. Chandrasekaran, and N. Papernot, "Entangled watermarks as a defense against model extraction," in *30th USENIX security symposium (USENIX Security 21)*, 2021, pp. 1937–1954.
[49] S. Szyller, B. G. Atli, S. Marchal, and N. Asokan, "Dawn: Dynamic adversarial watermarking of neural networks," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 4417–4425.
[50] Z. Wang, Y. Wu, and H. Huang, "Defense against model extraction attack by bayesian active watermarking," in *Forty-first International Conference on Machine Learning*, 2024.
[51] T. Orekondy, B. Schiele, and M. Fritz, "Prediction poisoning: Towards defenses against dnn model stealing attacks," *arXiv preprint arXiv:1906.10908*, 2019.
[52] M. Tang, A. Dai, L. DiValentin, A. Ding, A. Hass, N. Z. Gong, Y. Chen *et al.*, "{ModelGuard}:{Information-Theoretic} defense against model extraction attacks," in *33rd USENIX Security Symposium (USENIX Security 24)*, 2024, pp. 5305–5322.
[53] D.-D. Wu, C. Fu, W. Wu, W. Xia, X. Zhang, J. Zhou, and M.-L. Zhang, "Efficient model stealing defense with noise transition matrix," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 24 305–24 315.
[54] H. Fang, T. Zhang, T. Zhuang, J. Kong, K. Gao, B. Chen, L. Liang, S.-T. Xia, and K. Xu, "Towards distillation-resistant large language models: An information-theoretic perspective," *arXiv preprint arXiv:2602.03396*, 2026.
[55] G. Pereyra, G. Tucker, J. Chorowski, Ł. Kaiser, and G. Hinton, "Regularizing neural networks by penalizing confident output distributions," *arXiv preprint arXiv:1701.06548*, 2017.
[56] S. A. Balanya, J. Maronas, and D. Ramos, "Adaptive temperature scaling for robust calibration of deep neural networks," *Neural Computing and Applications*, vol. 36, no. 14, pp. 8073–8095, 2024.
[57] T. Joy, F. Pinto, S.-N. Lim, P. H. Torr, and P. K. Dokania, "Sample-dependent adaptive temperature scaling for improved calibration," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 12, 2023, pp. 14 919–14 926.
[58] A. S. Mozafari, H. S. Gomes, W. Leão, S. Janny, and C. Gagné, "Attended temperature scaling: a practical approach for calibrating deep neural networks," *arXiv preprint arXiv:1810.11586*, 2018.
[59] X.-C. Li, W.-S. Fan, S. Song, Y. Li, S. Yunfeng, D.-C. Zhan *et al.*, "Asymmetric temperature scaling makes larger networks teach well again," *Advances in neural information processing systems*, vol. 35, pp. 3830–3842, 2022.
[60] R. I. of Technology, "Research computing services," 2026. [Online]. Available: https://www.rit.edu/researchcomputing/

Khawaja Abaid Ullah received the B.S. degree in Computer Science from the University of Narowal, Pakistan in 2021, and the M.S. degree in Artificial Intelligence from the Rochester Institute of Technology (RIT), Rochester, NY, USA, in 2026, as a Fulbright Scholar. He is currently pursuing the Ph.D. degree in Electrical and Computer Engineering at RIT, serving as a Graduate Research Assistant in the Khojasteh Autonomous Systems (KAS) Lab. His research interests include deep learning, robotics, and autonomous systems.

Mohammad Javad Khojasteh (Member, IEEE) received the Ph.D. and M.Sc. degrees in electrical and computer engineering from the University of California, San Diego in 2019 and 2017, respectively. He is a Gleason Endowed Assistant Professor with the Rochester Institute of Technology (RIT). Before joining RIT, he held postdoctoral positions at Marine Physical Laboratory (MPL) at Scripps Institution of Oceanography (SIO), Department of Mechanical Engineering and Laboratory for Information and Decision Systems (LIDS) at Massachusetts Institute of Technology (MIT), and Center for Autonomous Systems and Technologies (CAST) at California Institute of Technology (Caltech), where he worked with Team CoSTAR as visitor at NASA Jet Propulsion Laboratory. He received the Gleason Chair in 2024 and the 2025 Provost's Learning Innovation Grant at RIT. Dr. Khojasteh's publications, co-authored with colleagues and students, have received awards, including Tammy L. Blair Student Paper Award (second place) from the International Society of Information Fusion. He is an Associate Editor for IEEE Signal Processing Letters.

APPENDIX

A. Notation

Table 3 summarizes the symbols used throughout the paper, grouped by role.

B. Calibration of the finite-difference step ε

Following the ADS paper [24], the penalty (2) approximates the exact directional derivative $\langle \nabla_\theta \log p(y \mid \mathbf{x}; \theta_{\mathcal{P}}), g \rangle$, which we compute directly as a Jacobian–vector product (JVP) on a reference batch. We select ε by sweeping over a log-grid and scoring the per-coordinate magnitude ratio and angular agreement between (2) and the JVP. Too-small ε loses the signal to floating-point cancellation; too-large ε leaves the linear regime (Fig. 15). The selected steps, $\varepsilon = 3 \times 10^{-3}$ on both CIFAR datasets and $\varepsilon = 10^{-2}$ on Tiny-ImageNet, each achieve a magnitude ratio of ≈ 1.00 under our GPU (TF32) inference stack; the calibration is hardware-arithmetic-specific and should be re-run if the serving stack changes numerics.

C. Sensitivity to the margin floor m

The guarantee is m-independent: recomputing the full serving pipeline over $m \in \{0.01, 0.05, 0.1, 0.2\}$ at $\lambda \in \{1, 2\}$, the defended teacher's top-1 drop is exactly 0.00 pp at every setting on both datasets, as Proposition 1 requires. What m modulates is only how much poison the budget passes, and mildly: at $\lambda = 2$ on CIFAR-100 the mean effective strength moves from 1.028 ($m = 0.01$) to 1.009 ($m = 0.2$) and

TABLE 3. Summary of notation.

Symbol	Description
K, Δ^{K-1}	Number of classes; probability simplex
$\mathcal{T}, \mathcal{S}, \mathcal{P}$	Teacher, student (attacker), proxy models
$\mathcal{P}_+, \mathcal{P}_-$	Frozen proxy clones at $\theta_{\mathcal{P}} \pm \varepsilon g$
$\mathbf{z}$, $\sigma(\cdot)$	Logit vector; softmax
t, $p_{\max}$	Served top class; top-class probability
gap_j	Margin $z_t - z_j$ to rival j
ℓ, g	Proxy holdout loss (per-sample mean); its gradient
$\widehat{\Delta}(y \mid \mathbf{x})$	Antidistillation penalty (2)
c_j, $c(\mathbf{x})$	Rival gain rate $\widehat{\Delta}_j - \widehat{\Delta}_t$; contrast (4)
λ, $\lambda(\mathbf{x})$	Requested strength; per-input budget (8)
$\lambda^*(\mathbf{x})$	Flip threshold (5)
$F(\lambda)$, λ_c	Flip-threshold CDF (6); transition threshold
m	Served-margin floor
τ, T	Served temperature (Secs. V-D–VI)
q	Budget-enforcement probability (Sec. IX-A)
$\lambda_{\min}$	Minimum served strength (Sec. IX-B)
ε	Finite-difference step
$\hat{\mathbf{p}}$	Served (defended) probability vector
$D_{KL}(\cdot\|\cdot)$	Kullback–Leibler divergence

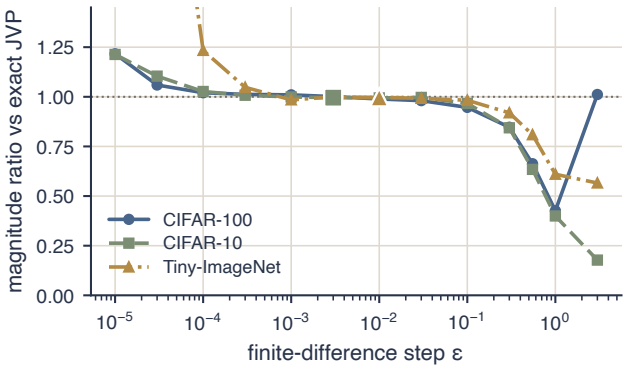
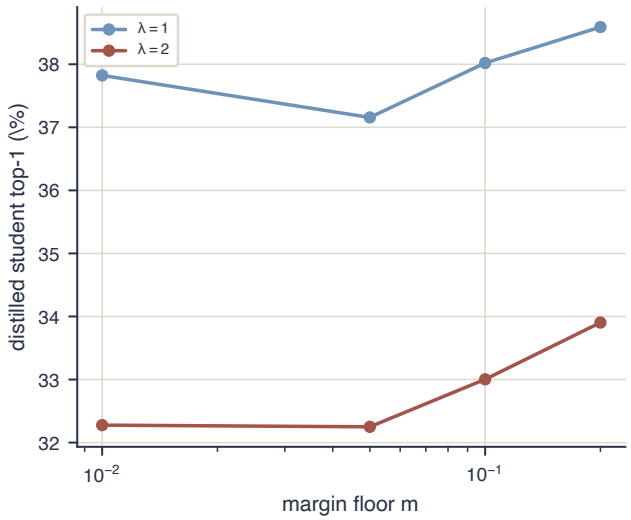
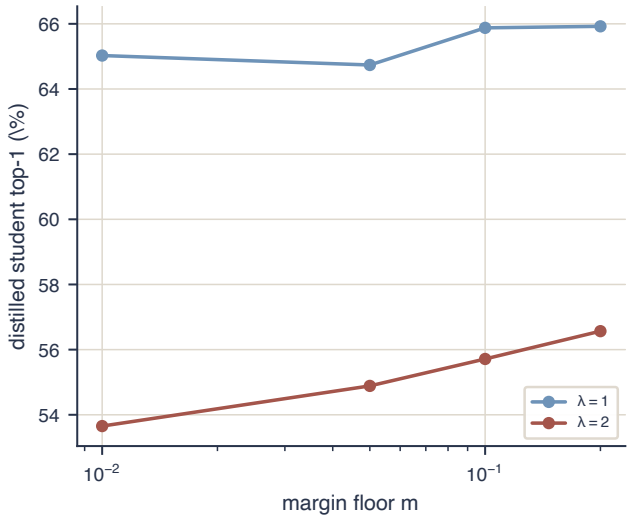


FIGURE 15. Calibration of the finite-difference step. Magnitude ratio between (2) and the exact JVP as a function of ε (log axis); the enlarged marker on each curve is the selected step. Below the plateau, floating-point cancellation destroys the signal; above it, the finite difference leaves the linear regime.

the served poison-KL from 0.842 to 0.757 (≈10% across a 20× range of m); CIFAR-10 behaves identically (0.702 → 0.627). At $\lambda = 1$ the spread is smaller still. Conclusions are therefore insensitive to m; we fix $m = 0.05$ as a mid-range served-margin service level. (Rows whose original margin is below m receive zero poison and keep their original margin, per the $\min(m, \text{original})$ guarantee—hence the minimum achieved served margin can lie below m.) A student-side confirmation (distilled students at $m \in \{0.01, 0.05, 0.1, 0.2\}$, $\lambda \in \{1, 2\}$, three seeds per cell) shows the same insensitivity



(a) CIFAR-100



(b) CIFAR-10

FIGURE 16. Student-side sensitivity to the margin floor m ($\lambda \in \{1, 2\}$, 3 seeds). Distilled-student top-1 varies by at most 1.7 pp (CIFAR-100) and 2.9 pp (CIFAR-10) over the $20\times$ range $m \in [0.01, 0.2]$, against a poison-induced degradation of 10–27 pp at these strengths; the defended teacher's top-1 drop is exactly zero at every cell.

where it matters (Fig. 16): student top-1 varies by at most 1.7 pp across the 20× range of m on CIFAR-100 ($\lambda = 2$: 32.3% at $m = 0.01$ versus 33.9% at $m = 0.2$) and by at most 2.9 pp on CIFAR-10 (53.7% versus 56.6%), against a poison-induced degradation of 10–27 pp at these λ. Larger floors admit marginally less poison, as expected, and the defended teacher's top-1 drop remains exactly zero at every cell.

D. Softening operators used in the ablation

The ablation of Sec. VIII-D uses a gated, argmax-preserving softener defined in log-odds space. Let $G = z_t -$

TABLE 4. Training recipes (SGD, momentum 0.9, Nesterov, weight decay 10^{-4}, cosine annealing throughout).

	Model	Epochs	LR	Batch
C-100	Teacher (ResNet-110)	200	0.1	256
	Proxy (ResNet-20)	50	0.1	256
	Student (ResNet-20)	50	0.1	512
C-10	Teacher (ResNet-56)	200	0.1	256
	Proxy (ResNet-20)	50	0.1	256
	Student (ResNet-20)	50	0.1	512
Tiny-IN	Teacher (ResNet-50†)	40	0.02	128
	Proxy (ResNet-18)	60	0.1	128
	Student (ResNet-18)	50	0.1	256

†ImageNet-pretrained, fine-tuned end to end at 64×64.

$\mathrm{logsumexp}_{j \neq t}\, z_j$, which equals $\mathrm{logit}(p_{\max})$ exactly. Given a gate p_g with $G_g = \mathrm{logit}(p_g)$, inputs with $G \leq G_g$ are untouched; over-confident inputs are mapped to the served confidence $G' = G_g$ by lowering the top logit only ($z_t \leftarrow G' + \mathrm{logsumexp}_{j \neq t}\, z_j$), leaving all non-target structure intact for the dark-knowledge channel. We sweep $p_g \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$; the constant-temperature comparison is Sec. VI (E-CT). The operator is inspired in spirit by asymmetric temperature scaling [59], which applies separate temperatures to the correct and wrong classes so that wrong-class discriminability survives softening; ours is the gated, serving-side limiting case in which the wrong-class logits are left exactly unchanged. Additionally, where asymmetric temperature scaling selects a fixed pair of temperatures by grid search over a candidate space, our operator adapts the effective softening per query.

E. Reproducibility

Table 4 reports the full training recipes. All models are trained with SGD (momentum 0.9, Nesterov, weight decay 10^{-4}) under cosine annealing; teachers and proxies use a 5-epoch linear warmup, plain cross-entropy, and *no* label smoothing. Teachers and proxies train on the augmented training split (random crop with 4-pixel padding at 32×32, 8-pixel at 64×64, plus horizontal flip); the Tiny-ImageNet teacher is an ImageNet-pretrained ResNet-50 fine-tuned end to end with the stem adapted to 64×64 (first-conv stride $2 \to 1$, max-pool removed, fresh 200-way head). The clone gradient g is computed on a held-out 10% of the training split. The attacker's student trains on the cached (image, served-vector) pairs with no augmentation. Every figure is generated by a script that emits a sibling metrics JSON with the underlying numbers, and aggregates for every (variant, λ, seed) cell are archived. Code, configuration files, and exact command lines: https://github.com/the-kas-lab/ADS-C.