DISTILLING BAYESIAN UNCERTAINTY INTO A SINGLE FORWARD PASS

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep ensembles give reliable Bayesian uncertainty estimates, but they need K forward passes for every input. We treat the uncertainty of a Bayesian teacher as a function of the input and train a single deterministic network, our student, to output it in one forward pass. Distilling such posterior quantities has been done before, but the targets are Monte Carlo estimates from K members: they change when the ensemble is retrained, and what a student learns from these noisy targets has not been analysed. We model the student's uncertainty head as a regression on the noisy teacher uncertainty, allowing correlated noise and a biased teacher, and give an exact condition under which the student is closer to the ideal uncertainty than its teacher. When the teacher's bias and the part of its noise that does not average out over members are small, the best regularisation strength decreases as K grows. A controlled simulation and ensembles trained from scratch follow the predicted trend in K. To learn the uncertainty function, our student model first learns the teacher's predictions; a small head on the frozen backbone then learns the teacher's uncertainty on clean, corrupted, mixed and masked inputs, with a log-space regression loss and a ranking loss. Any teacher that provides this uncertainty can be used. On image classification, segmentation and text, our student stays close to the teacher's accuracy at the cost of one forward pass. Scored against ground-truth labels on image classification, it has the lowest or tied-lowest selective-prediction risk among single-pass methods that do not distil the teacher, is competitive with other distillation methods, and detects distribution shift without labels as well as the teacher. Its out-of-distribution detection is close to the teacher's.

1 INTRODUCTION

Uncertainty quantification (UQ) is needed wherever a model's errors are costly, for example in medical diagnosis or autonomous driving. Bayesian deep learning offers many ways to approximate the posterior and read its epistemic uncertainty (EU): deep ensembles, MC Dropout, Laplace and SWAG approximations, and single-model evidential or distance-aware methods. Deep ensembles (Lakshminarayanan et al., 2017) are widely used as the reference: their predictions and uncertainty are among the most reliable under distribution shift (Snoek et al., 2019; Gustafsson et al., 2020), whereas the EU of single-model methods has been shown to be relative rather than absolute, fragile, or poorly calibrated under shift (Bengs et al., 2022; Kopetzki et al., 2021; Postels et al., 2022; Jürgens et al., 2024; Shen et al., 2024). An ensemble, however, needs K forward passes per input at test time.

We want both: the predictions and uncertainty of the best available Bayesian method, at the cost of one forward pass. Our view is that uncertainty is a function of the input, $x \mapsto u^\star(x)$, defined by that method; we do not build a new approximation of the posterior but learn this function. A teacher, here a deep ensemble, evaluates it offline on training and auxiliary inputs. Our student, a single deterministic network, learns the teacher's predictions and uncertainty in two stages and returns both in one forward pass, as in amortized inference (Kingma & Welling, 2022; Gershman & Goodman, 2014). Any method that yields $u^\star$ can be the teacher. Distilling posterior quantities is not new (Balan et al., 2015; Vadera et al., 2020; Malinin et al., 2019). Two questions have received less attention. First, the targets are noisy: a K-member estimate of uncertainty changes when the ensemble is retrained. What does regression on such targets do, and when can a student trained on

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

them be closer to the ideal target than the teacher that produced them? Second, is the student's score useful when it is scored against ground-truth labels rather than against the teacher?

Contributions. (1) A conditional analysis of distillation from noisy targets (§4): we model the uncertainty head as kernel ridge regression and describe the teacher only by its mean, bias and noise covariance, with no independence or unbiasedness assumed. We give an exact condition under which the student has lower expected error than its teacher, and show that, when the teacher's bias and the noise that does not average out over members are small, the best regularisation strength decreases as K grows. A controlled simulation and ensembles trained from scratch follow the predicted trend in K (§5.4). **(2) A recipe for learning the uncertainty function** (§3): Stage 1 distils the teacher's predictions; Stage 2 trains a small head on the frozen backbone to output the teacher's MI, with a log-space regression loss and a ranking loss, on an auxiliary set of clean, corrupted, mixed and masked inputs. Each part is ablated. **(3) An evaluation that separates fidelity from utility** (§5): fidelity is the agreement between the student's score and the teacher's uncertainty; utility is how well the score works against ground-truth labels. Since the analysis shows that a student need not be closer to the ideal target than its teacher at large K, we judge the score mainly by its utility. On image classification, segmentation, and text, our student stays within 2% of teacher accuracy at 4.9–$16\times$ higher throughput. On all four image-classification benchmarks it has the lowest or tied-lowest AURC (selective-prediction risk) among single-pass methods that do not distil the teacher, and it has the highest power among eight label-free shift monitors on CIFAR-10-C. Its OOD detection is close to the teacher's as well.

2 Background and Related Work

Bayesian uncertainty. For input x, label y, training data $\mathcal{D}$ and parameters θ with posterior $p(\theta \mid \mathcal{D})$, the posterior predictive $p(y \mid x, \mathcal{D}) = \int p(y \mid x, \theta)\, p(\theta \mid \mathcal{D})\, d\theta$ splits total uncertainty (TU) into aleatoric (AU) and epistemic (EU) parts (Kendall & Gal, 2017; Depeweg et al., 2018):

$$\mathrm{TU} = H[p(y \mid x, \mathcal{D})], \quad \mathrm{AU} = \mathbb{E}_{p(\theta \mid \mathcal{D})} H[p(y \mid x, \theta)], \quad \mathrm{EU} = \mathrm{TU} - \mathrm{AU} = I[y; \theta \mid x, \mathcal{D}], \quad (1)$$

with H the entropy and I the mutual information. Each is a scalar function of the input,

$$u^\star(x) = \mathcal{F}\big(p(\theta \mid \mathcal{D}),\, p(y \mid x, \theta)\big), \quad (2)$$

which is what we distil. Deep ensembles (Lakshminarayanan et al., 2017) estimate it with K members and remain a strong reference (Snoek et al., 2019), at K times the inference cost.

Cheaper and single-model uncertainty. Sampling methods (MC Dropout (Gal & Ghahramani, 2016), Laplace (Ritter et al., 2018), SWAG (Maddox et al., 2019), Packed Ensembles (Laurent et al., 2023)) still need several passes or sub-networks per input. Evidential methods (Sensoy et al., 2018; Amini et al., 2019; Charpentier et al., 2020) predict the parameters of a Dirichlet or Normal-Inverse-Gamma distribution from labels alone; their epistemic estimates can differ from Bayesian ones (Bengs et al., 2022; Shen et al., 2024). Distance-aware methods (DUQ (van Amersfoort et al., 2020), SNGP (Liu et al., 2020a), DDU (Mukhoti et al., 2023), DUE (van Amersfoort et al., 2022)) read uncertainty from feature-space geometry and constrain the backbone, e.g. with spectral normalisation. Post-hoc scores such as Energy (Liu et al., 2020b) and ViM (Wang et al., 2022) are designed as OOD detectors; we also score them on selective prediction. All of these are ways to approximate Bayesian uncertainty, and single-model estimates in particular have known weaknesses in what they can represent (Kopetzki et al., 2021; Postels et al., 2022; Jürgens et al., 2024). We do not propose another approximation: we take the uncertainty of a reference method as given and learn the function $u^\star$ of Eq. (2); the methods above serve as reference points in our evaluation.

Distilling ensembles and posterior quantities. Knowledge distillation (KD) (Hinton et al., 2015) trains a student on a teacher's outputs. Bayesian Dark Knowledge (Balan et al., 2015) distils the posterior predictive of an SGLD teacher online, and Vadera & Marlin (2019) study how robust this is to posterior uncertainty. Generalized Posterior Expectation Distillation (GPED) (Vadera et al., 2020) extends it to posterior expectations of arbitrary functions, including entropy-based uncertainty; it is the closest prior work, and the abstraction of distilling a posterior expectation into one network is theirs. Ensemble distribution distillation (Malinin et al., 2019; Fathullah & Gales, 2022) and Hydra (Tran et al., 2020) keep the members' diversity in a Dirichlet output or in multiple heads, so that uncertainty can be computed from one model. Shen et al. (2021) distil an uncertainty-aware

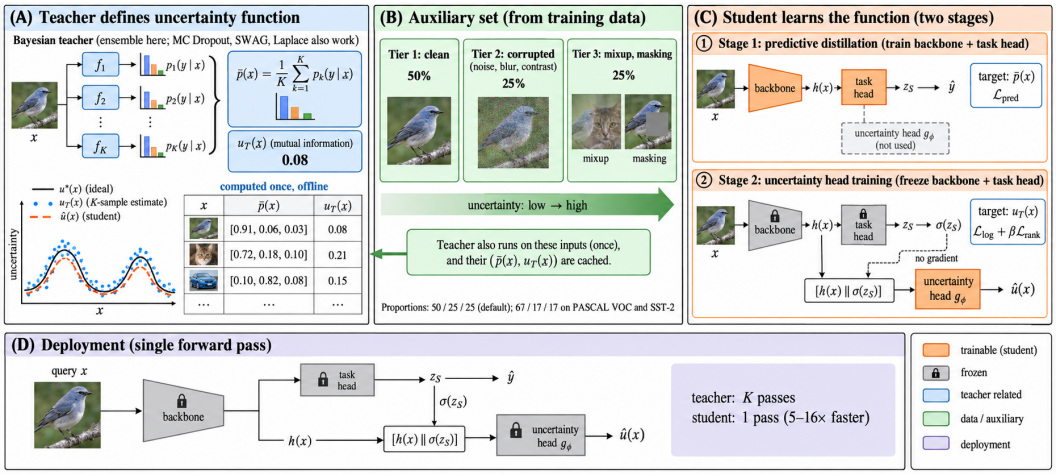

Figure 1: **Overview. (A)** A Bayesian teacher (here an ensemble) is run once offline on the training and auxiliary inputs; its averaged prediction $\bar{p}(x)$ and uncertainty $u_T(x)$ are cached. u_T is a K-sample estimate of $u^\star$ (§4). **(B)** The auxiliary set spans low to high uncertainty: clean, corrupted, and mixup or masked inputs. **(C)** Stage 1 distils $\bar{p}$ into the student; Stage 2 freezes the backbone and task head and fits the uncertainty head g_ϕ to u_T. **(D)** At deployment one forward pass gives $\hat{y}$ and $\hat{u}(x)$.

distribution for real-time vision, and Holder & Shafique (2021) and DUDES (Landgraf et al., 2023) distil ensemble uncertainty maps for segmentation. Contextual similarity distillation (Zanger et al., 2025) targets the variance of the infinite-width NTK Gaussian process of an ensemble. None of these analyses the finite-K Monte Carlo noise in the distillation targets. Relative to these methods, our student learns the uncertainty function itself with a head on a frozen backbone, and our analysis and evaluation address the two questions of §1; App. B.1 compares them in a table.

3 Method

Our student outputs a prediction $\hat{y}$ and a scalar uncertainty $\hat{u}(x)$ that approximates the teacher's. The pipeline has four parts (Figure 1): offline teacher targets (§3.1), two-stage training (§3.2), a loss suited to the range of the targets (§3.3), and an auxiliary training set (§3.4).

3.1 Teacher Targets

In the main experiments the teacher is a deep ensemble with member predictive distributions $p_k(y \mid x)$, $k = 1, \ldots, K$, and the target is its epistemic uncertainty, the mutual information (MI) between the label and the member index (Depeweg et al., 2018):

$$u_T(x) = H\Big[\tfrac{1}{K}\textstyle\sum_{k=1}^{K} p_k(y \mid x)\Big] - \tfrac{1}{K}\textstyle\sum_{k=1}^{K} H\big[p_k(y \mid x)\big]. \tag{3}$$

This is a post-hoc statistic of the members' softmax outputs. It is computed once for every training and auxiliary input and cached (exceptions in App. C.1), and the student is deployed without the teacher. u_T is a K-sample estimate of the functional $u^\star$ of Eq. (2), and it changes when the ensemble is retrained; §4 models this variation. The pipeline needs only u_T, so another Bayesian teacher can take the ensemble's place (App. B.10), and other scalar functionals, such as a regression ensemble's epistemic variance, can be distilled the same way (§5.7).

3.2 Two-Stage Training

Prediction and uncertainty compete for the same backbone features, and training them jointly lowers every metric (Table 5; App. A.7 gives an information-theoretic analysis), so we train in two stages.

Stage 1: predictive distillation. The student backbone and task head are trained with standard knowledge distillation (Hinton et al., 2015):

$$\mathcal{L}_{\text{pred}} = (1-\alpha)\,\text{CE}(y,\hat{y}) + \alpha\,\tau^2\,\text{KL}\big(\sigma(z_T/\tau)\,\big\|\,\sigma(z_S/\tau)\big), \tag{4}$$

where σ is the softmax, z_S the student logits, $z_T = \log \bar{p}$ the log of the teacher's averaged predictive distribution $\bar{p} = K^{-1}\sum_k p_k$, τ the temperature and α a mixing weight.

Stage 2: uncertainty head. The backbone and task head are frozen. A head g_ϕ with one hidden layer receives the penultimate features and the detached student softmax, $[\mathbf{h}(x)\,\|\,\sigma(z_S)]$, and is trained on the cached targets. On Tiny ImageNet the input also includes two feature-geometry values, the largest cosine similarity to and the smallest distance from the training class means; they help the head fit the teacher's MI (Pearson 0.812 vs. 0.769 without them) and raise mean OOD AUROC by 0.012 (App. C.1). For segmentation the head is a 1×1-convolution head on the encoder features, without the softmax input. Its output passes through a Softplus, so $g_\phi \geq 0$; on MNIST the output is linear, clamped at zero at test time, and trained with plain rather than log-space MSE (App. C.1).

3.3 Loss

MI targets are close to zero on most clean inputs and heavily right-skewed on ambiguous or shifted inputs. We combine a regression term in log space (Eigen et al., 2014), which keeps small and large targets on a comparable scale, with a pairwise ranking term (Burges et al., 2005), which enforces their order:

$$\begin{aligned}\mathcal{L}_{\log} &= \frac{1}{m}\sum_{i=1}^{m}\Big(\log\big(1+g_\phi(x_i)\big) - \log\big(1+u_T(x_i)\big)\Big)^2,\\ \mathcal{L}_{\text{rank}} &= \frac{1}{|\mathcal{P}|}\sum_{(i,j)\in\mathcal{P}}\max\big(0,\ \delta - (g_\phi(x_i) - g_\phi(x_j))\big),\end{aligned} \tag{5}$$

where $\mathcal{P}$ contains the pairs with $u_T(x_i) > u_T(x_j)$ among 256 pairs drawn at random from a batch, and $\delta = 0.05$ is a margin. The Stage-2 loss is $\mathcal{L}_{\text{unc}} = \mathcal{L}_{\log} + \beta\,\mathcal{L}_{\text{rank}}$ with $\beta = 1$ on every classification and segmentation benchmark. The log term is the part analysed in §4. The ranking term defines no regression target and is outside that analysis; its role is shown by ablation (§5.6); per-benchmark variants: App. C.1.

3.4 Auxiliary Training Set

On clean training inputs the teacher's MI is nearly constant and close to zero, so a head trained only on them sees little variation. Stage 2 therefore uses three tiers built from the training set: **Tier 1**, clean inputs (low uncertainty); **Tier 2**, corrupted inputs, e.g. Gaussian noise, blur and reduced contrast at one fixed severity each on CIFAR-10 (intermediate); **Tier 3**, mixup of two randomly paired inputs with the mixing weight cycled over a fixed grid, which creates class ambiguity, and inputs with large occlusions, which remove information (high); Tiny ImageNet adds further Tier-3 types. The tiers make up $50/25/25\%$ of the Stage-2 set (VOC and SST-2: $67/17/17\%$). Segmentation and text use task-specific Tier-3 inputs (region masking and mixup; token masking). No OOD evaluation set is used for training. Details are in App. C.1 and Algorithm 1.

4 Analysis: When Can the Student Remove Teacher Noise?

The teacher's target u_T (Eq. (3)) is a K-sample Monte Carlo estimate, so it changes when the ensemble is retrained. The uncertainty head g_ϕ of Stage 2 is fit to u_T by regularised regression and therefore smooths it. This section asks when the smoothing removes more of the teacher's noise than of its signal. We study a surrogate for the head: population kernel ridge regression (KRR) on the log-space target of Eq. (5). The surrogate covers the log-regression term (App. A.1). The results are conditions. §5.4 tests them in a controlled simulation, where every term is known, and checks their observable consequences on trained models.

4.1 SETUP

Let P be the Stage-2 training distribution (the three-tier auxiliary mixture of §3.4) and $\langle f, g\rangle = \mathbb{E}_P[fg]$ the inner product of $L^2(P)$. In this section u denotes the log-space target $\log(1 + \mathrm{EU})$; rank-based metrics are unchanged by this monotone map, and, after clipping the student to the range of MI, L^2 errors transfer back to the EU scale with a factor of at most $1 + \log C$ for C classes (App. A.2). Let k be a Mercer kernel with integral operator T, eigenvalues $t_1 \geq t_2 \geq \cdots > 0$ and orthonormal eigenfunctions ϕ_j in $L^2(P)$, and let $\mathcal{H}_k$ be its reproducing kernel Hilbert space (RKHS). The surrogate student is

$$u_S = \arg\min_{f \in \mathcal{H}_k} \|f - u_T\|^2_{L^2(P)} + \lambda \|f\|^2_{\mathcal{H}_k} = R_\lambda u_T, \qquad R_\lambda = T(T + \lambda I)^{-1}, \tag{6}$$

so in the eigenbasis the student keeps a fraction $\rho_j = t_j/(t_j + \lambda) \in (0, 1)$ of each teacher coefficient $a_j = \langle u_T, \phi_j\rangle$: directions with large t_j (smooth, low-frequency) are kept and directions with small t_j are shrunk (Engl et al., 1996; Mobahi et al., 2020).

The randomness of interest is retraining the K-member teacher. We describe it only through its first two moments, with no independence across inputs and no unbiasedness:

$$m_K = \mathbb{E}[u_T], \qquad \beta_K = m_K - u^\star, \qquad C_K = \mathrm{Cov}(u_T), \tag{7}$$

where $u^\star$ is the ideal clean uncertainty of Eq. (2), β_K is the teacher's bias, and C_K is the covariance operator of the teacher's error on $L^2(P)$. We assume neither independence nor unbiasedness because neither holds in practice: the same K members produce u_T at every input, so the errors at different inputs are correlated (C_K is not diagonal); and across independently retrained ensembles the error decays as $1/K$ towards a non-zero floor (§5.4).

4.2 MOMENTS OF THE STUDENT SCORE

Theorem 1 (Moments under P; proof in App. A.3). *Let $\Delta = \mathbb{E}_P[u_T] - \mathbb{E}_P[u_S] = \langle \mathbf{1}, (I - R_\lambda)u_T\rangle$ be the mean shift. For every realisation of u_T: (a) $\mathbb{E}_P[u_T^2] - \mathbb{E}_P[u_S^2] = \langle u_T, (I - R_\lambda^2)u_T\rangle = \sum_j (1-\rho_j^2)a_j^2 \geq 0$, non-decreasing in λ; (b) $\mathrm{Var}_P(u_T) - \mathrm{Var}_P(u_S) = \sum_j (1-\rho_j^2)a_j^2 - \Delta\,(\mathbb{E}_P[u_T] + \mathbb{E}_P[u_S])$, so the variance contracts if and only if $\Delta\,(\mathbb{E}_P[u_T] + \mathbb{E}_P[u_S]) < \sum_j (1 - \rho_j^2)a_j^2$; (c) $|\Delta| \leq \|(I - R_\lambda)u_T\|$, and this bound is non-decreasing in λ.*

The second moment always contracts; the variance contracts only when the mean shift is small. The downstream results below need variance contraction, so we treat it as a premise and check it.

4.3 FINITE-K COMPARISON WITH THE TEACHER

Theorem 2 (Student versus teacher; proof in App. A.4). *With m_K, β_K, C_K as in Eq. (7) and a kernel that does not depend on the teacher draw,*

$$\mathbb{E}\|u_S - u^\star\|^2 - \mathbb{E}\|u_T - u^\star\|^2 = \|(I - R_\lambda)m_K\|^2 - 2\langle \beta_K, (I - R_\lambda)m_K\rangle - \mathrm{tr}\big[(I - R_\lambda^2)C_K\big]. \tag{8}$$

Hence the student has strictly lower expected error if and only if $\mathrm{tr}[(I - R_\lambda^2)C_K] + 2\langle \beta_K, (I - R_\lambda)m_K\rangle > \|(I - R_\lambda)m_K\|^2$. In the eigenbasis, $\mathrm{tr}[(I - R_\lambda^2)C_K] = \sum_j (1 - \rho_j^2)\,\mathrm{Var}(a_j)$.

In this condition, $\mathrm{tr}[(I - R_\lambda^2)C_K]$ is the teacher noise that the smoothing removes, $\|(I - R_\lambda)m_K\|^2$ is the signal that the smoothing removes, and the middle term measures how the teacher's bias β_K aligns with that removed signal; it can have either sign. Three consequences follow, with t_j the kernel eigenvalues and λ the regularisation strength. (i) The gain is large when much of the teacher's noise lies in directions with small t_j while m_K lies in directions with large t_j. Nothing requires the noise to be independent across inputs; correlated noise enters through C_K. (ii) If $C_K = C_0 + C_1/K$, with C_1/K the part of the noise that averages out over members and C_0 the part that does not, the removable noise tends to $\mathrm{tr}[(I - R_\lambda^2)C_0]$ as $K \to \infty$ while the signal loss at fixed λ does not shrink. So a fixed λ can win at every K only if the floor C_0 or the bias term makes up for its signal loss. If $C_0 = 0$ and $\beta_K = 0$, no fixed λ wins for all K, and a second-order expansion gives a best λ_K that decreases roughly as $1/K$ (Remark 1); for small C_0 and β_K the best λ_K still decreases with K. (iii) Teacher bias in the directions the filter keeps ($\rho_j \approx 1$) passes to the student almost unchanged, so distillation cannot correct a teacher that is systematically wrong there. Unlike James–Stein shrinkage (James & Stein, 1961), this is not a dominance result.

4.4 Downstream Use

OOD detection and batch monitoring use a score through its mean gap between two distributions and its spread within each. App. A.5 shows: if the student's variance is no larger than the teacher's on the distributions involved and the mean shifts of Theorem 1 cancel up to ε between them, then the student's standardised ID/OOD separation and the power of a mean-shift test are at least the teacher's, up to $O(\varepsilon)$ (Corollary 1, Theorem 3). Both premises can be checked with unlabelled data by comparing the two scores. On every CIFAR-10 and CIFAR-100 ID/OOD pair, where we use these results, the variance premise and the bound of Corollary 1 hold (§5.4); on MNIST the teacher's uncertainty on ID inputs is close to zero and the premise does not hold.

5 Experiments

We evaluate prediction and cost (C1), uncertainty against ground truth (C2), the conditions of §4 (C3) and deployment (C4).

5.1 Experimental Protocol

Benchmarks and teachers. Table 1 lists the six benchmarks: MNIST (LeCun et al., 1998), CIFAR-10 and CIFAR-100 (Krizhevsky & Hinton, 2009), Tiny ImageNet (Le & Yang, 2015), PASCAL VOC 2012 segmentation (Everingham et al., 2010), and SST-2 sentiment classification (Socher et al., 2013). Every teacher is a deep ensemble whose members differ in seed and in training hyperparameters drawn at random per member (which ones varies by benchmark; App. C.1): 4-layer CNNs on MNIST and ResNet-18 (He et al., 2015) on CIFAR-10 and CIFAR-100, all trained from scratch, and LoRA (Hu et al., 2022) fine-tunes of one pretrained DeiT-Small (Touvron et al., 2020), SegFormer-B2 (Xie et al., 2021) or BERT-base (Devlin et al., 2019) backbone on the other three. Our student uses the teacher's architecture, except on SST-2 where it is DistilBERT (Sanh et al., 2019). Fine-tuned members start from one pretrained solution, so we test the analysis of §4 only on the three ensembles trained from scratch. MNIST, CIFAR-10 and Tiny ImageNet are evaluated on 5 / 6 / 7 OOD sets (App. C.1), CIFAR-100 on the OpenOOD v1.5 protocol (Zhang et al., 2024), and CIFAR-10-C (Hendrycks & Dietterich, 2019) is used for the ablations and for monitoring.

Model selection and seeds. For each training run we keep one checkpoint (epoch), chosen by prediction accuracy or by label-free agreement with the teacher, in some runs on the evaluation data, and never by an uncertainty metric (App. B.8); the Tiny ImageNet Tier-2 inputs are corrupted test images. With a validation split held out from training instead, mean OOD AUROC and AURC change by at most 0.005 on CIFAR-10 and Tiny ImageNet, and the Tiny ImageNet Pearson correlation with the teacher drops from 0.842 to 0.814 (App. B.8). Over five Stage-2 seeds the standard deviation of mean OOD AUROC is 0.0006 (CIFAR-10) and 0.0038 (Tiny ImageNet). All CIFAR-100 models are selected on a held-out validation split, and CIFAR-100 student results average three Stage-2 seeds.

Baselines. Multi-pass: MC Dropout (Gal & Ghahramani, 2016), last-layer Laplace (LLLA) (Ritter et al., 2018), SGLD (Welling & Teh, 2011), SWAG (Maddox et al., 2019). Single-pass: EDL (Sensoy et al., 2018), SNGP (Liu et al., 2020a), DDU (Mukhoti et al., 2023), DUE (van Amersfoort et al., 2022), DEUP (Lahlou et al., 2023), and the post-hoc OOD scores Energy (Liu et al., 2020b) and ViM (Wang et al., 2022) computed on the backbone of one teacher member and on the student backbone. On CIFAR-100 we run MC Dropout, DDU, Energy and ViM. The distillation baselines EnD^2 (Malinin et al., 2019) and Hydra (Tran et al., 2020) are compared on CIFAR-10 and, for EnD^2, CIFAR-100 (§5.3). Baseline settings are in App. D.8.

Metrics. *Fidelity* is the Pearson and Spearman correlation between a score and the teacher's MI; in the main text we compare it only across distillation methods (all methods: App. D.1). *Ground-truth utility* is measured for every method with its own score: the area under the risk–coverage curve when abstaining in decreasing order of uncertainty (AURC, lower is better), the AUROC of separating the method's errors from its correct predictions (misclassification AUROC), OOD AUROC averaged over the OOD sets of a benchmark, and the squared kernel calibration error (SKCE) of the predictive distribution. A difference counts as a tie when its 95% paired-bootstrap interval contains zero.

Table 1: **Teachers, accuracy, and deployment speed-up.** Accuracy (for VOC, mIoU computed per image and averaged; dataset-level mIoU in App. F.1) of the K-member teacher and of our single-pass student, top-1 agreement between them, and the ratio of student to teacher throughput, all measured with one protocol on one GPU (App. D.7).

Benchmark	Teacher members	K	Teacher ↑	Student ↑	Δ	Agree.	Speed-up ↑
MNIST	4-layer CNN	16	99.54%	99.51%	−0.03	99.8%	15.5×
CIFAR-10	ResNet-18	16	96.96%	95.99%	−0.97	97.5%	16.0×
CIFAR-100	ResNet-18	5	80.78%	78.99%	−1.79	85.7%	5.0×
Tiny ImageNet	DeiT-S + LoRA	16	88.13%	87.50%	−0.63	92.9%	16.1×
VOC 2012 (mIoU)	SegFormer-B2 + LoRA	5	0.802	0.806	+0.004	–	4.9×
SST-2	BERT-base + LoRA	5	91.51%	92.09%	+0.58	94.6%	9.9×

Table 2: **Ground-truth utility.** AURC and misclassification AUROC on each ID test set; OOD AUROC averaged over all OOD sets of the benchmark. "Best other" is the best method other than ours, the teacher and the distillation methods (Table 3 and §5.3): single-pass for the first two metrics, any method for OOD. The teacher needs K passes. Full grids: App. B.2.

	AURC ↓			Miscls. AUROC ↑			Mean OOD AUROC ↑		
	Ours	Best other	Teacher	Ours	Best other	Teacher	Ours	Best other	Teacher
MNIST	1.5×10^{-4}	1.7×10^{-4} SNGP	1.1×10^{-4}	0.973	0.980 EDL	0.979	0.991	0.997 MCD	0.994
CIFAR-10	0.0034	0.0048 ViM‡	0.0023	0.942	0.946 EDL	0.947	0.901	0.905 ViM‡	0.903
CIFAR-100	0.0593	0.0639 Energy‡	0.0539	0.853	0.847 Energy‡	0.846	0.807	0.813 Energy‡	0.796
Tiny ImageNet	0.0243	0.0315 EDL	0.0224	0.888	0.895 EDL	0.892	0.860	0.913 ViM§	0.897

MCD: MC Dropout (16 passes). Post-hoc scores on the student's backbone (‡) or on one teacher member's (§). CIFAR-100: K=5, OpenOOD v1.5 sets, baselines as listed in §5.1.

5.2 C1: Prediction and Cost

Our student is within 1% of the teacher on five benchmarks and 1.8% below it on CIFAR-100, and runs at the speed of one member (twice that on SST-2, with DistilBERT). Training the full CIFAR-10 pipeline (K=16) takes 0.52 GPU-hours on one RTX 4090, which the faster deployment recovers after 4.1M samples (App. B.6).

5.3 C2: Uncertainty Quality Against Ground Truth

Our student has lower AURC than every single-pass baseline in Table 2 on CIFAR-10, CIFAR-100 and Tiny ImageNet, beyond noise, and ties SNGP on MNIST. On misclassification AUROC it ties the best single-pass method (EDL) on MNIST, CIFAR-10 and Tiny ImageNet; on CIFAR-100 it reaches 0.853, against 0.847 for the best baseline in Table 2. On clean test data its predictive calibration matches the teacher's: the student–teacher SKCE difference is a tie on MNIST, CIFAR-10 and Tiny ImageNet (App. B.2). Under CIFAR-10-C corruptions the teacher is better calibrated (App. D.6).

On the ensembles trained from scratch, OOD detection is close to the teacher's. In mean OOD AUROC, our student ties the teacher on CIFAR-10 (0.901 vs. 0.903) and is above it on CIFAR-100 (0.807 vs. 0.796); on the near-OOD sets of the OpenOOD v1.5 CIFAR-10 protocol both reach 0.930 (App. B.12). Post-hoc OOD scores (Energy, ViM) computed on our student's backbone are 0.004 (CIFAR-10) and 0.006 (CIFAR-100) higher in mean OOD AUROC and have higher AURC than our uncertainty head (Table 2). On Tiny ImageNet, where the fine-tuned teacher (0.897) is itself below ViM (0.913), our student reaches 0.860.

Fidelity and other distillation methods. A retrained teacher does not reproduce the original uncertainty exactly, so the agreement between two independently trained teachers is a natural scale for fidelity: our student's Spearman correlation with its teacher is within noise of it on CIFAR-10 (0.790 vs. 0.802) and above it on CIFAR-100 (App. B.3). Table 3 compares our student with EnD^2, which fits a Dirichlet to the members' predictions, and Hydra, which keeps one head per member on a shared body. On CIFAR-10 our student has the lowest AURC, the highest Pearson fidelity and the highest mean OOD AUROC (by 0.004); Hydra is about 0.1 points more accurate. On CIFAR-100 an EnD^2 head on our student's backbone has lower AURC (0.054 vs. 0.059), higher misclassification

Table 3: **Distillation methods on CIFAR-10** (K=16 teacher; ResNet-18 body, Hydra with 16 heads). Ours is the model of Tables 1–2; the other heads and baselines are selected on a held-out split (ours with the same selection: App. B.8). The first two rows share one backbone. Fidelity: Pearson / Spearman correlation with the teacher's MI.

Method	Acc. ↑	AURC ↓	Mean OOD ↑	Fidelity ↑
Ours	95.99%	0.0034	0.901	0.815 / 0.799
EnD^2 head on our backbone	95.72%	0.0038	0.897	0.684 / 0.773
EnD^2 (Malinin et al., 2019), full network	94.84%	0.0046	0.883	0.713 / 0.800
Hydra (Tran et al., 2020), MLP heads	96.13%	0.0060	0.832	0.734 / 0.543
Hydra, residual-block heads	96.08%	0.0052	0.853	0.681 / 0.598

Table 4: **Checks of the analysis in §4**, at three levels (Surr.: surrogate, a controlled KRR simulation; Direct: measured on trained networks; Down.: premises of the downstream results). The reference is built from ensembles retrained from scratch with other seeds. Details in App. B.4.

Level	Check	Result
Surr.	Student/teacher error, best λ	0.01–0.02 / 0.42–0.52 / 0.98–1.00 for noise where the filter shrinks / flat / where the signal lies
Direct	Teacher error vs. reference	MNIST, CIFAR-10: decays as $1/K$ to a floor ($R^2 = 0.98$–0.99)
Direct	Student/teacher error vs. ref.	CIFAR-10: 0.83, 1.04, 1.81 at K=2, 5, 16; higher at the largest K than at K=2 on all three
Direct	Best λ vs. K (ridge)	CIFAR-10, trained features: 1, 1, 1, 0.3, 0.1 at K=2, 3, 5, 10, 16
Direct	Thm. 2 condition, β_K=0	CIFAR-10: holds for $\lambda/\lambda_{\max} \leq 10^{-3}$ (grid from 10^{-6})
Down.	Premises and bounds	Cor. 1: variance contracts, Eq. (17) holds on all 12 CIFAR-10/100 ID/OOD pairs. Thm. 3(c): $\sigma_S \leq \sigma_T$ in 75/75, standardised shift $\geq$ teacher's in 67/75 scenarios

AUROC (0.875 vs. 0.853) and higher fidelity, and our student has higher mean OOD AUROC (0.807 vs. 0.798); all three differences are beyond noise. Other Bayesian teachers also work (App. B.10).

5.4 C3: Do the Conditions of §4 Hold?

Table 4 summarises the checks. In the controlled simulation, where every term is known, the outcome depends on where the noise lies, as Theorem 2 predicts, and the best λ falls with K (Remark 1).

On trained networks, the ideal target $u^\star$ is unknown. We therefore retrain further ensembles from scratch with other seeds and use their MI as an independent reference for $u^\star$. Against this reference, the teacher's error decays as $1/K$ to a floor (Remark 2), and the ratio of our student's squared error to the teacher's is higher at the largest K than at K=2 on MNIST, CIFAR-10 and CIFAR-100 (App. B.4). This is what Theorem 2 predicts when the teacher's noise shrinks with K and the student's smoothing does not. The trained head selects weight decay 0 or near 0 at every K but still smooths the target (in aggregate it contracts the teacher's score), so it acts like a fixed $\lambda > 0$ in Eq. (8).

5.5 C4: Deployment

Label-free monitoring. After deployment, inputs arrive in unlabelled batches, and some batches come from a shifted distribution on which accuracy drops; a monitor must flag these batches from a score alone. For each batch, a one-sided test compares the mean score with its mean on clean reference data (Theorem 3; false-positive rate 0.05), and power is the fraction of shifted batches flagged. On CIFAR-10-C (75 corruption–severity scenarios), our student's score has the highest average power of eight monitors (0.932), above the K=16 teacher (0.925) and the 12-signal Suitability Filter of Pouget et al. (2025) (0.910). Power grows with the standardised shift (batch-mean shift over the score's standard deviation), and our student's standardised shift is at least the teacher's in 67 of 75 scenarios (App. F.3.1).

Table 5: **Ablations on CIFAR-10 (K=16).** Top: Stage-2 variants on the same frozen Stage-1 backbone. Bottom: joint training of both stages. Mean OOD AUROC over six OOD sets; CIFAR-10-C AUROC over 19 corruptions × 5 severities.

Variant	Acc. ↑	Pearson ↑	Spearman ↑	Mean OOD ↑	CIFAR-10-C ↑
Ours	95.99%	0.815	0.799	0.901	0.719
Tier 1 only	95.99%	0.800	0.807	0.884	0.698
No Tier 3	95.99%	0.814	0.807	0.905	0.715
Ranking loss only	95.99%	0.704	0.818	0.901	0.716
Log-space MSE only	95.99%	0.824	0.650	0.907	–
Plain MSE (no log, no ranking)	95.99%	0.814	0.644	0.909	0.716
Joint training	95.07%	0.646	0.627	0.853	0.686

Cascade routing. Our student answers every input and sends the inputs it is most uncertain about to the teacher; the question is how few inputs must be sent to reach the teacher's accuracy. Sending 10% reaches it on CIFAR-10 (96.99% vs. 96.96%) at 2.6 instead of 16 passes per input (App. F.3.2).

5.6 Ablations

Joint training lowers every metric. The two loss terms trade off: without the ranking term Spearman falls from 0.799 to 0.650, ranking alone lowers Pearson to 0.704, and plain MSE has the highest mean OOD AUROC (0.909). Adding Tier 2 (corrupted inputs) to Tier 1 (clean inputs) raises mean OOD AUROC by 0.021; Tier 3 (mixup and masking) changes mean OOD and CIFAR-10-C AUROC by less than 0.005. The ranking weight β sets the same trade-off (from β=0 to 5, Pearson falls from 0.828 to 0.778 and Spearman rises from 0.659 to 0.807), and a smaller teacher lowers fidelity (Pearson 0.63 at K=3 vs. 0.81 at K=16) but not mean OOD AUROC (0.897–0.907; App. E.2).

5.7 Segmentation, Text, and Regression

On PASCAL VOC our per-pixel student reaches 0.806 mIoU (teacher 0.802) at 4.9× the speed (App. F.1); on SST-2 it matches the teacher's accuracy, with higher OOD AUROC and lower AURC than the teacher (App. F.2); the same two stages also distil regression ensembles (App. B.7).

6 Conclusion and Limitations

In one forward pass, our student approximates an ensemble's predictions and epistemic uncertainty: on six vision and text benchmarks it stays close to the teacher's accuracy, and on image classification its score is competitive with single-pass methods. Uncertainty estimates are usually judged by selective prediction and OOD detection; we also use ours for label-free monitoring of a deployed model, where on CIFAR-10-C its average power is above that of the ensemble it was distilled from and of the Suitability Filter, a monitor designed for this task, at the cost of one forward pass. Whether it also removes the teacher's Monte Carlo noise depends on where the noise lies, on the teacher's bias and on the regularisation strength; on the three ensembles trained from scratch, its squared error relative to the teacher's is higher at large K than at small K, as the analysis predicts when its smoothing does not shrink with K.

Limitations. (1) *Teacher dependence.* Our student copies its teacher's weaknesses; a bias shared by all ensembles trained with one recipe can be neither removed nor measured with independent pools. (2) *Scope of the analysis.* The kernel surrogate omits the ranking loss, finite width and sample size, and feature learning; it is tested in aggregate on three ensembles trained from scratch. (3) *Absolute scale.* High fidelity on ID inputs does not guarantee a meaningful scale on OOD inputs: under repeated distillation the OOD scores drift and can change sign (App. F.4). (4) *Cost.* Running a large teacher on all training and auxiliary inputs is expensive; offline training pays off after about 4M deployed samples (App. B.6).

AI USE STATEMENT

We used large language models (ChatGPT and Claude) in preparing this paper to restructure and edit the text, check derivations (especially the biased-teacher case of Theorem 2 and the expansion in Remark 1), and assist with experimental design. AI agents were also used to monitor long-running experiments, debug experimental configurations, suggest subsequent hyperparameter settings based on intermediate results, and summarize experimental outcomes. Figure 1 was generated with ChatGPT based on specifications provided by the authors. The authors derived the theoretical results independently, checked the derivations against AI suggestions, compared all reported numbers with the experiment logs, and confirmed each reference. The authors take full responsibility for the content of this paper, including all text, claims, results, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

App. A states all assumptions and gives complete proofs. App. C lists the architectures, hyperparameters, auxiliary-set construction and sizes, and baseline configurations; Algorithm 1 summarises the training procedure; App. B.6 gives the timed cost of each stage. An anonymised repository (`https://anonymous.4open.science/r/uncertainty-distillation-ICLR46724`) contains the code for target generation, the auxiliary set, both training stages and the evaluation, with end-to-end scripts and cached teacher targets for CIFAR-10 and MNIST; the code for the baselines and for the analyses in App. B will be released with the final version.

ETHICS STATEMENT

This work uses public benchmark datasets and involves no human subjects. A distilled uncertainty score inherits the failures of its teacher and can give false confidence under shifts the teacher handles poorly; we document these failure cases (§5.3, §6) so that users can judge where the score can be relied on.

REFERENCES

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. *CoRR*, abs/1910.02600, 2019. URL http://arxiv.org/abs/1910.02600.

Anoop Korattikara Balan, Vivek Rathod, Kevin P. Murphy, and Max Welling. Bayesian dark knowledge. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 3438–3446, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/af3303f852abeccd793068486a391626-Abstract.html.

Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. Pitfalls of epistemic uncertainty quantification through loss minimisation. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/bc1d640f841f752c689aae20b31198c1-Abstract-Conference.html.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pp. 89–96, 2005.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational mathematics*, 7(3):331–368, 2007.

Bertrand Charpentier, Daniel Zügner, and Stephan Günnemann. Posterior network: Uncertainty estimation without OOD samples via density-based pseudo-counts. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien

Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/0eac690d7059a8de4b48e90f14510391-Abstract.html.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Jacob Cohen. *Statistical power analysis for the behavioral sciences*. routledge, 2013.

Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International conference on machine learning*, pp. 1184–1193. PMLR, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pp. 2366–2374, 2014. URL https://proceedings.neurips.cc/paper/2014/hash/7bccfde7714a1ebadf06c5f4cea752c1-Abstract.html.

Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

Yassir Fathullah and Mark JF Gales. Self-distribution distillation: efficient uncertainty estimation. In *Uncertainty in Artificial Intelligence*, pp. 663–673. PMLR, 2022.

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

Samuel Gershman and Noah D. Goodman. Amortized inference in probabilistic reasoning. In Paul Bello, Marcello Guarini, Marjorie McShane, and Brian Scassellati (eds.), *Proceedings of the 36th Annual Meeting of the Cognitive Science Society, CogSci 2014, Quebec City, Canada, July 23-26, 2014*. cognitivesciencesociety.org, 2014. URL https://escholarship.org/uc/item/34j1h7k5.

David M. Green and John A. Swets. *Signal detection theory and psychophysics*. Wiley, New York, 1966.

Fredrik K. Gustafsson, Martin Danelljan, and Thomas B. Schön. Evaluating scalable bayesian deep learning methods for robust computer vision. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020*, pp. 1289–1298. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPRW50498.2020.00167. URL https://openaccess.thecvf.com/content_CVPRW_2020/html/w20/Gustafsson_Evaluating_Scalable_Bayesian_Deep_Learning_Methods_for_Robust_Computer_Vision_CVPRW_2020_paper.html.

Bharath Hariharan, Pablo Arbeláez, Lubomir D. Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In Dimitris N. Metaxas, Long Quan, Alberto Sanfeliu, and Luc Van Gool (eds.), *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011*, pp. 991–998. IEEE Computer Society, 2011. doi: 10.1109/ICCV.2011.6126343. URL https://doi.org/10.1109/ICCV.2011.6126343.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. URL http://arxiv.org/abs/1503.02531.

Christopher J. Holder and Muhammad Shafique. Efficient uncertainty estimation in semantic segmentation via distillation. In *IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2021, Montreal, QC, Canada, October 11-17, 2021*, pp. 3080–3087. IEEE, 2021. doi: 10.1109/ICCVW54120.2021.00343. URL https://doi.org/10.1109/ICCVW54120.2021.00343.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Arthur Jacot, Clément Hongler, and Franck Gabriel. Neural tangent kernel: Convergence and generalization in neural networks. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 8580–8589, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/5a4be1fa34e62bb8a6ec6b91d2462f5a-Abstract.html.

William James and Charles Stein. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 361–379. University of California Press, 1961.

Mira Jürgens, Nis Meinert, Viktor Bengs, Eyke Hüllermeier, and Willem Waegeman. Is epistemic uncertainty faithfully represented by evidential deep learning methods? In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22624–22642. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/juergens24a.html.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5574–5584, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/2650d6089a6d640c5e85b2b88265dc2b-Abstract.html.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. URL https://arxiv.org/abs/1312.6114.

Anna-Kathrin Kopetzki, Bertrand Charpentier, Daniel Zügner, Sandhya Giri, and Stephan Günnemann. Evaluating robustness of predictive uncertainty estimation: Are dirichlet-based models reliable? In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of*

Machine Learning Research, pp. 5707–5718. PMLR, 2021. URL http://proceedings.mlr.press/v139/kopetzki21a.html.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Ananya Kumar, Percy Liang, and Tengyu Ma. Verified uncertainty calibration. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 3787–3798, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/f8c0c968632845cd133308b1a494967f-Abstract.html.

Salem Lahlou, Moksh Jain, Hadi Nekoei, Victor Butoi, Paul Bertin, Jarrid Rector-Brooks, Maksym Korablyov, and Yoshua Bengio. DEUP: direct epistemic uncertainty prediction. *Trans. Mach. Learn. Res.*, 2023, 2023. URL https://openreview.net/forum?id=eGLdVRvvfQ.

Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 6402–6413, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.

Steven Landgraf, Kira Wursthorn, Markus Hillemann, and Markus Ulrich. DUDES: deep uncertainty distillation using ensembles for semantic segmentation. *CoRR*, abs/2303.09843, 2023. doi: 10.48550/ARXIV.2303.09843. URL https://doi.org/10.48550/arXiv.2303.09843.

Olivier Laurent, Adrien Lafage, Enzo Tartaglione, Geoffrey Daniel, Jean-Marc Martinez, Andrei Bursuc, and Gianni Franchi. Packed ensembles for efficient uncertainty estimation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=XXTyv1zD9zD.

Y. Le and X. Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791. URL https://doi.org/10.1109/5.726791.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 7167–7177, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/abdeb6f575ac5c6676b747bca8d09cc2-Abstract.html.

E.L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer, Cham, 2022. ISBN 978-3-030-70577-0. doi: 10.1007/978-3-030-70578-7. URL https://link.springer.com/book/10.1007/978-3-030-70578-7.

Jeremiah Z. Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax-Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020a. URL https://proceedings.neurips.cc/paper/2020/hash/543e83748234f7cbab21aa0ade66565f-Abstract.html.

Jeremiah Zhe Liu, Shreyas Padhy, Jie Ren, Zi Lin, Yeming Wen, Ghassen Jerfel, Zachary Nado, Jasper Snoek, Dustin Tran, and Balaji Lakshminarayanan. A simple approach to improve single-model deep uncertainty via distance-awareness. *J. Mach. Learn. Res.*, 24:42:1–42:63, 2023. URL https://jmlr.org/papers/v24/22-0479.html.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020b. URL https://proceedings.neurips.cc/paper/2020/hash/f5496252609c43eb8a3d147ab9b9c006-Abstract.html.

Wesley J. Maddox, Pavel Izmailov, Timur Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 13132–13143, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/118921efba23fc329e6560b27861f0c2-Abstract.html.

Andrey Malinin and Mark J. F. Gales. Uncertainty estimation in autoregressive structured prediction. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=jN5y-zb5Q7m.

Andrey Malinin, Bruno Mlodozeniec, and Mark J. F. Gales. Ensemble distribution distillation, 2019. URL http://arxiv.org/abs/1905.00076.

Hossein Mobahi, Mehrdad Farajtabar, and Peter L. Bartlett. Self-distillation amplifies regularization in hilbert space, 2020. URL https://arxiv.org/abs/2002.05715.

Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip H. S. Torr, and Yarin Gal. Deep deterministic uncertainty: A new simple baseline. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 24384–24394. IEEE, 2023. doi: 10.1109/CVPR52729.2023.02336. URL https://doi.org/10.1109/CVPR52729.2023.02336.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Ng. Reading digits in natural images with unsupervised feature learning. *NIPS*, 01 2011.

Janis Postels, Mattia Segù, Tao Sun, Luca Daniel Sieber, Luc Van Gool, Fisher Yu, and Federico Tombari. On the practicality of deterministic epistemic uncertainty. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 17870–17909. PMLR, 2022. URL https://proceedings.mlr.press/v162/postels22a.html.

Angéline Pouget, Mohammad Yaghini, Stephan Rabanser, and Nicolas Papernot. Suitability filter: A statistical framework for classifier evaluation in real-world deployment settings. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/pouget25a.html.

Hippolyt Ritter, Aleksandar Botev, and David Barber. A scalable laplace approximation for neural networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=Skdvd2xAZ.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019. URL http://arxiv.org/abs/1910.01108.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 525–536, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/432aca3a1e345e339f35a30c8f65edce-Abstract.html.

Murat Sensoy, Lance M. Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 3183–3193, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/a981f2b708044d6fb4a71a1463242520-Abstract.html.

Maohao Shen, Jongha Jon Ryu, Soumya Ghosh, Yuheng Bu, Prasanna Sattigeri, Subhro Das, and Gregory W. Wornell. Are uncertainty quantification capabilities of evidential deep learning a mirage? In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/c3177be226ee12e34d6ba3b5e6fe6a5b-Abstract-Conference.html.

Yichen Shen, Zhilu Zhang, Mert R. Sabuncu, and Lin Sun. Real-time uncertainty estimation in computer vision via uncertainty-aware distribution distillation. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pp. 707–716. IEEE, 2021. doi: 10.1109/WACV48630.2021.00075. URL https://doi.org/10.1109/WACV48630.2021.00075.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information, 2017. URL http://arxiv.org/abs/1703.00810.

Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26:153–172, 08 2007. doi: 10.1007/s00365-006-0659-y.

Jasper Snoek, Yaniv Ovadia, Emily Fertig, Balaji Lakshminarayanan, Sebastian Nowozin, D. Sculley, Joshua V. Dillon, Jie Ren, and Zachary Nado. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 13969–13980, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 1631–1642. ACL, 2013. doi: 10.18653/V1/D13-1170. URL https://doi.org/10.18653/v1/d13-1170.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, and Nathan Srebro. The implicit bias of gradient descent on separable data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=r1q7n9gAb.

Trevor Standley, Amir Zamir, Dawn Chen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9120–9132. PMLR, 2020. URL http://proceedings.mlr.press/v119/standley20a.html.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Information Science and Statistics. Springer, New York, 2008.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *CoRR*, abs/1910.10699, 2019. URL http://arxiv.org/abs/1910.10699.

Naftali Tishby, Fernando C. N. Pereira, and William Bialek. The information bottleneck method, 2000. URL http://arxiv.org/abs/physics/0004057.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention, 2020. URL https://arxiv.org/abs/2012.12877.

Linh Tran, Bastiaan S. Veeling, Kevin Roth, Jakub Swiatkowski, Joshua V. Dillon, Jasper Snoek, Stephan Mandt, Tim Salimans, Sebastian Nowozin, and Rodolphe Jenatton. Hydra: Preserving ensemble diversity for model distillation. *CoRR*, abs/2001.04694, 2020. URL https://arxiv.org/abs/2001.04694.

Meet P. Vadera and Benjamin M. Marlin. Assessing the robustness of bayesian dark knowledge to posterior uncertainty. *CoRR*, abs/1906.01724, 2019. URL http://arxiv.org/abs/1906.01724.

Meet P. Vadera, Brian Jalaian, and Benjamin M. Marlin. Generalized bayesian posterior expectation distillation for deep neural networks. In Ryan P. Adams and Vibhav Gogate (eds.), *Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence, UAI 2020, virtual online, August 3-6, 2020*, volume 124 of *Proceedings of Machine Learning Research*, pp. 719–728. AUAI Press, 2020. URL http://proceedings.mlr.press/v124/vadera20a.html.

Joost van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9690–9700. PMLR, 2020. URL http://proceedings.mlr.press/v119/van-amersfoort20a.html.

Joost van Amersfoort, Lewis Smith, Andrew Jesson, Oscar Key, and Yarin Gal. On feature collapse and deep kernel learning for single forward pass uncertainty, 2022. URL https://arxiv.org/abs/2102.11409.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.

Gal Vardi. On the implicit bias in deep-learning algorithms. *Commun. ACM*, 66(6):86–93, 2023. doi: 10.1145/3571070. URL https://doi.org/10.1145/3571070.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 4911–4920. IEEE, 2022. doi: 10.1109/CVPR52688.2022.00487. URL https://doi.org/10.1109/CVPR52688.2022.00487.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. In Lise Getoor and Tobias Scheffer (eds.), *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pp. 681–688. Omnipress, 2011. URL https://icml.cc/2011/papers/398_icmlpaper.pdf.

David Widmann, Fredrik Lindsten, and Dave Zachariah. Calibration tests in multi-class classification: A unifying framework. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 12236–12246, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/1c336b8080f82bcc2cd2499b4c57261d-Abstract.html.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017. URL http://arxiv.org/abs/1708.07747.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, José M. Álvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 12077–12090, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/64f1f27bf1b4ec22924fd0acb550c235-Abstract.html.

Greg Yang and Edward J. Hu. Tensor programs IV: feature learning in infinite-width neural networks. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11727–11737. PMLR, 2021. URL http://proceedings.mlr.press/v139/yang21c.html.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836, 2020.

Moritz A. Zanger, Pascal R. van der Vaart, Wendelin Böhmer, and Matthijs T. J. Spaan. Contextual similarity distillation: Ensemble uncertainties with a single model. *CoRR*, abs/2503.11339, 2025. doi: 10.48550/ARXIV.2503.11339. URL https://doi.org/10.48550/arXiv.2503.11339.

Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. Openood v1.5: Enhanced benchmark for out-of-distribution detection. *J. Data-centric Mach. Learn. Res.*, 2:(3):1–32, 2024. URL https://data.mlr.press/assets/pdf/v02-3.pdf.

APPENDIX ROADMAP

App. A: assumptions, scope of the kernel surrogate, and proofs of Theorems 1–2, Corollary 1 and Theorem 3; App. A.7 gives an information-theoretic account of two-stage training. **App. B**: positioning against prior distillation work, the ground-truth evaluation, all checks of the analysis (simulation, independent pools, moment audit, eNTK), the teacher-diversity ablation, timed cost, and the regression study. **App. C**: hyperparameters, auxiliary-set construction (Algorithm 1), baseline mechanisms, and win/tie/loss counts. **App. D**: per-method tables for fidelity, OOD detection, selective prediction, calibration, throughput, and the baseline scope matrix. **App. E**: full ablation sweeps and diagnostic figures. **App. F**: segmentation, SST-2, monitoring, cascade routing, and repeated self-distillation.

A ANALYSIS: ASSUMPTIONS AND PROOFS

This appendix states the setting of §4, lists what the surrogate does not cover, and proves the results. Notation follows §4.1: $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ are those of $L^2(P)$, $R_\lambda = T(T+\lambda I)^{-1}$ with eigenvalues $\rho_j = t_j/(t_j+\lambda)$, and $a_j = \langle u_T, \phi_j\rangle$.

A.1 SETTING AND SCOPE OF THE SURROGATE

Assumption 1 (Square integrability)**.** *$u^\star, u_T, u_S \in L^2(P)$ and $\mathbb{E}\|u_T\|^2 < \infty$ over teacher draws. When a result involves another distribution Q, the functions also belong to $L^2(Q)$.*

Assumption 2 (Kernel surrogate)**.** *$\mathcal{X}$ is compact and k is a continuous, universal Mercer kernel, so that $\{\phi_j\}$ is a complete orthonormal basis of $L^2(P)$ (Steinwart & Christmann, 2008). The head is modelled by the population KRR estimator of Eq.* (6)*, which exists and is unique for every $\lambda > 0$ (Caponnetto & De Vito, 2007; Smale & Zhou, 2007). For Theorem 2, k does not depend on the teacher draw.*

Assumption 3 (Teacher moments)**.** *The teacher draw is a retraining of the K members. Its mean m_K, bias $\beta_K = m_K - u^\star$ and covariance operator $C_K = \mathbb{E}[(u_T - m_K)\otimes(u_T - m_K)]$ are as in Eq.* (7)*; C_K is trace class. No independence across inputs and no unbiasedness is assumed.*

What the surrogate does not cover. (1) *The ranking term.* $\mathcal{L}_{\mathrm{rank}}$ uses only the teacher's ordering of pairs (its margin acts on differences of the student's scores) and defines no regression target, so no statement here applies to it; its effect is measured by ablation (Table 5). (2) *Finite width and feature learning.* Wide networks trained with small steps behave like kernel regression with the neural tangent kernel (Jacot et al., 2018), but the trained head is not in that regime (Yang & Hu, 2021). In the head's empirical NTK basis, 52.3% of components are amplified rather than shrunk (App. B.4), so coordinate-wise agreement with $\rho_j \in (0,1)$ is not expected; only aggregate statements are tested. (3) *Finite samples.* The empirical KRR estimator contracts the target at the training points; at new points it need not. (4) *Dependence of the kernel on the teacher.* The head's features come from a backbone distilled from the teacher's logits (Stage 1), so in practice the kernel is not independent of the teacher draw, as Theorem 2 assumes. (5) *Distribution.* P is the Stage-2 training mixture. On a test or OOD distribution $Q \neq P$ the eigenfunctions are not orthonormal and contraction is not guaranteed (Eq. (18)).

A.2 TRANSFER FROM THE LOG-SPACE TARGET

The regression term of Eq. (5) fits $v = \log(1+u)$. The map $u \mapsto \log(1+u)$ is a strictly increasing bijection of $[0,\infty)$, so the analysis applies verbatim to v_T and v_S, and both losses have the same minimiser as $\lambda \to 0$. Rank-based quantities (Spearman correlation, AUROC, AURC, the ordering used for routing) are identical on the two scales. For errors, mutual information is bounded by $\log C$ for C classes; clipping v_S to $[0, \log(1+\log C)]$ can only reduce its error, and on this range $|e^v - e^{v'}| \le (1+\log C)\,|v - v'|$, so

$$\|u_S - u_T\| \le (1+\log C)\,\|v_S - v_T\|, \qquad \|u_S - u^\star\| \le (1+\log C)\,\|v_S - v^\star\|. \tag{9}$$

The factor is 3.3 for C=10 and 6.3 for C=200. In the controlled simulation (App. B.4) the L2 and log-space students behave alike for all three noise spectra.

A.3 Proof of Theorem 1

Proof. By completeness, $u_T = \sum_j a_j \phi_j$ and $u_S = R_\lambda u_T = \sum_j \rho_j a_j \phi_j$.

(a) Orthonormality gives $\mathbb{E}_P[u_T^2] = \sum_j a_j^2$ and $\mathbb{E}_P[u_S^2] = \sum_j \rho_j^2 a_j^2$, hence

$$\mathbb{E}_P[u_T^2] - \mathbb{E}_P[u_S^2] = \langle u_T, (I - R_\lambda^2) u_T \rangle = \sum_j (1 - \rho_j^2)\, a_j^2 \geq 0. \tag{10}$$

Since $\partial \rho_j / \partial \lambda = -t_j/(t_j + \lambda)^2 < 0$, every factor $1 - \rho_j^2$ is increasing in λ.

(b) Writing $\mathrm{Var}_P(u) = \mathbb{E}_P[u^2] - (\mathbb{E}_P[u])^2$ and $(\mathbb{E}_P u_T)^2 - (\mathbb{E}_P u_S)^2 = \Delta\,(\mathbb{E}_P u_T + \mathbb{E}_P u_S)$,

$$\mathrm{Var}_P(u_T) - \mathrm{Var}_P(u_S) = \sum_j (1 - \rho_j^2)\, a_j^2 - \Delta\,(\mathbb{E}_P[u_T] + \mathbb{E}_P[u_S]), \tag{11}$$

which gives the stated if-and-only-if condition. The second term need not be small: with non-negative scores both factors of the product are typically positive, and nothing in Assumptions 1–3 bounds Δ relative to the first sum.

(c) $\mathbb{E}_P[f] = \langle \mathbf{1}, f \rangle$ and $\|\mathbf{1}\| = 1$ because P is a probability measure, so by Cauchy–Schwarz $|\Delta| = |\langle \mathbf{1}, (I - R_\lambda) u_T \rangle| \leq \|(I - R_\lambda) u_T\| = (\sum_j (1 - \rho_j)^2 a_j^2)^{1/2}$, and each $1 - \rho_j$ is increasing in λ. The realised $|\Delta| = |\sum_j (1 - \rho_j) a_j \mathbb{E}_P[\phi_j]|$ need not be monotone, because its summands can cancel. □

Relative to the target, the student's mean error is $\mathbb{E}_P[u_S] - \mathbb{E}_P[u^\star] = -\Delta + (\mathbb{E}_P[u_T] - \mathbb{E}_P[u^\star])$; averaged over teacher draws the second term is $\langle \mathbf{1}, \beta_K \rangle$, the teacher's mean bias.

A.4 Proof of Theorem 2

Proof. Write $u_T = m_K + \varepsilon$ with $\mathbb{E}[\varepsilon] = 0$ and $\mathrm{Cov}(\varepsilon) = C_K$. Since R_λ is fixed, bounded and self-adjoint,

$$u_S - u^\star = R_\lambda m_K - u^\star + R_\lambda \varepsilon = \beta_K - (I - R_\lambda) m_K + R_\lambda \varepsilon. \tag{12}$$

Taking expectations over teacher draws, the cross term vanishes and $\mathbb{E}\|R_\lambda \varepsilon\|^2 = \mathrm{tr}(R_\lambda C_K R_\lambda) = \mathrm{tr}(R_\lambda^2 C_K)$ by cyclicity of the trace (C_K is trace class). Therefore

$$\mathbb{E}\|u_S - u^\star\|^2 = \|\beta_K\|^2 - 2\langle \beta_K, (I - R_\lambda) m_K \rangle + \|(I - R_\lambda) m_K\|^2 + \mathrm{tr}(R_\lambda^2 C_K), \tag{13}$$

$$\mathbb{E}\|u_T - u^\star\|^2 = \|\beta_K\|^2 + \mathrm{tr}(C_K). \tag{14}$$

Subtracting gives Eq. (8). Since $I - R_\lambda^2$ is diagonal in $\{\phi_j\}$, $\mathrm{tr}[(I - R_\lambda^2) C_K] = \sum_j (1 - \rho_j^2) \langle \phi_j, C_K \phi_j \rangle = \sum_j (1 - \rho_j^2) \mathrm{Var}(a_j)$; no diagonality of C_K is needed. □

In coordinates, with $m_j = \langle m_K, \phi_j \rangle$, $\beta_j = \langle \beta_K, \phi_j \rangle$ and $s_j = \mathrm{Var}(a_j)$, the student is better if and only if

$$\sum_j (1 - \rho_j)^2 m_j^2 - 2 \sum_j (1 - \rho_j)\, \beta_j m_j < \sum_j (1 - \rho_j^2)\, s_j. \tag{15}$$

Because s_j scales as $1/K$ for a Monte Carlo average and the other two terms do not, a fixed $\lambda > 0$ satisfies the condition for every K only if the bias term outweighs the signal loss or part of the noise does not average out.

Remark 1 (Small λ and the scaling of $\lambda_K^\star$)**.** *Let $D(\lambda)$ denote the left side of Eq.* (8) *(student minus teacher). Using $1 - \rho_j = \lambda/(t_j + \lambda)$,*

$$\frac{D(\lambda)}{\lambda} = \sum_j \frac{\lambda\, m_j^2}{(t_j + \lambda)^2} - 2 \sum_j \frac{\beta_j m_j}{t_j + \lambda} - \sum_j \frac{(2t_j + \lambda)\, s_j}{(t_j + \lambda)^2}. \tag{16}$$

If $m_K \in \mathcal{H}_k$ (i.e. $\sum_j m_j^2/t_j < \infty$) the first sum tends to 0 as $\lambda \to 0^+$ by dominated convergence ($\lambda/(t + \lambda)^2 \leq 1/(4t)$); the last sum increases monotonically to $2\sum_j s_j/t_j$. Hence, if $\sum_j |\beta_j m_j|/t_j < \infty$ and $\sum_j s_j/t_j + \sum_j \beta_j m_j/t_j > 0$, there is $\bar{\lambda} > 0$ such that the student is

strictly better for every $\lambda \in (0, \bar{\lambda})$. With $\beta_K = 0$ this holds whenever $C_K \neq 0$. A second-order expansion with $\beta_K = 0$ and $s_j = s_j^{(1)}/K$ gives, heuristically, $D(\lambda) \approx -2\lambda S_1/K + \lambda^2(M_2 + 3S_2/K)$ with $S_1 = \sum_j s_j^{(1)}/t_j$, $S_2 = \sum_j s_j^{(1)}/t_j^2$, $M_2 = \sum_j m_j^2/t_j^2$, and the minimiser of this approximation, $\lambda_K^\star = S_1/(KM_2 + 3S_2)$, i.e. $K\lambda_K^\star = S_1/(M_2 + 3S_2/K)$, which approaches S_1/M_2 from below as K grows (this needs m_K in the range of T, a stronger condition than $m_K \in \mathcal{H}_k$). In the controlled simulation with flat noise and the L2 loss the best λ is 7.1×10^{-4}, 2.8×10^{-4} and 7.9×10^{-5} at $K = 5, 16, 64$, i.e. $K\lambda_K^\star = 3.5, 4.5, 5.1$ ($\times10^{-3}$): increasing with K, as the expression predicts, so $\lambda_K^\star$ decays slightly more slowly than $1/K$. The expansion is not valid when much of the noise sits at very small t_j, which is the regime where large λ can help.

Remark 2 (What the error floor of independent pools measures)**.** *For independently retrained pools, §5.4 fits $\mathrm{MSE}(K) = a + b/K$ between a K-member teacher and a reference built from the other pools. With $C_K = C_0 + C_1/K$ and pools trained with the same recipe, $\mathbb{E}\|u_T^{(K)} - u_{\mathrm{ref}}\|^2 = \mathrm{tr}(C_0) + \mathrm{tr}(C_1)/K + \mathrm{tr}(C_{\mathrm{ref}}) + \|\beta_K - \beta_{\mathrm{ref}}\|^2$. The part of the bias shared by the recipe cancels; what remains is the difference between the biases of a K-member and a reference-size ensemble, which arises because the plug-in MI estimate depends on ensemble size. The intercept a therefore measures the variance that does not average out, the reference's own noise, and the large-K limit of this bias difference (its K-dependent part enters the fitted slope); it is not the teacher's bias β_K. Theorem 2 allows both $C_0 \neq 0$ and $\beta_K \neq 0$. A bias shared by every ensemble trained with the same recipe cannot be identified from such pools at all. With a fixed reference, $\mathrm{tr}(C_{\mathrm{ref}})$ can be estimated by splitting the reference in two, and $\mathrm{tr}(C_0)$ from half the variance of the difference between two full pools minus the member-level term; on CIFAR-10 they account for 0.76 and 0.49 of the intercept $a = 1.68$ ($\times10^{-3}$), leaving 0.43 for the bias difference and estimation error (App. B.4).*

A.5 Downstream results

Standardised separation. For an ID distribution P_0 and an OOD distribution P_1, let $\mu_k(u)$ and $V_k(u)$ be the mean and variance of a score u under P_k, and $\Gamma(u) = (\mu_1(u) - \mu_0(u))/\sqrt{V_0(u) + V_1(u)}$, a variant of Cohen's d (Cohen, 2013) and of d' in signal detection (Green & Swets, 1966).

Corollary 1 (Separation under explicit premises)**.** *Assume (i) the mean gap is preserved up to ε: $\mu_1(u_S) - \mu_0(u_S) \geq \mu_1(u_T) - \mu_0(u_T) - \varepsilon$, with $\mu_1(u_T) - \mu_0(u_T) > \varepsilon$; and (ii) $V_k(u_S) \leq V_k(u_T)$ for $k = 0, 1$. Then*

$$\Gamma(u_S) \;\geq\; \Gamma(u_T) - \frac{\varepsilon}{\sqrt{V_0(u_T) + V_1(u_T)}}. \tag{17}$$

If (ii) holds only for $k = 0$, then $\Gamma(u_S) \geq (\mu_1(u_T) - \mu_0(u_T) - \varepsilon)/\sqrt{V_0(u_T) + V_1(u_S)}$.

Proof. By (i) the numerator of $\Gamma(u_S)$ is at least $\mu_1(u_T) - \mu_0(u_T) - \varepsilon > 0$; by (ii) its denominator is at most $\sqrt{V_0(u_T) + V_1(u_T)}$. Dividing gives Eq. (17); the partial case is identical. □

Premise (i) holds when the mean shifts Δ of Theorem 1, evaluated under P_0 and P_1, nearly cancel. Premise (ii) is not implied by Theorem 1: that theorem concerns P and needs its mean condition, while under $Q \neq P$ the eigenfunctions have Gram matrix $G_{jk}^Q = \mathbb{E}_Q[\phi_j \phi_k] \neq \delta_{jk}$ and

$$\mathbb{E}_Q[u_T^2] - \mathbb{E}_Q[u_S^2] = \sum_{j,k} (1 - \rho_j \rho_k)\, a_j a_k\, G_{jk}^Q, \tag{18}$$

which can be negative. Both premises can be checked from unlabelled samples of P_0 and P_1 by comparing the two scores. The moment audit of App. B.4 does this for every ID and OOD split. Because (ii) needs both variances to contract, we count ID/OOD pairs: it holds for all 6 CIFAR-10 and all 6 CIFAR-100 pairs and for none of the 17 MNIST, Tiny ImageNet and SST-2 pairs (their ID split, or every split, does not contract).

Label-free monitoring. The monitor follows the Suitability Filter of Pouget et al. (2025) with a single score. After training, compute the reference profile of the score on a large reference set $\mathcal{D}_{\mathrm{ref}}$,

$$\mu_0 = \frac{1}{|\mathcal{D}_{\mathrm{ref}}|} \sum_{x \in \mathcal{D}_{\mathrm{ref}}} u(x), \qquad \sigma_0^2 = \frac{1}{|\mathcal{D}_{\mathrm{ref}}|} \sum_{x \in \mathcal{D}_{\mathrm{ref}}} (u(x) - \mu_0)^2, \tag{19}$$

treated as known because $|\mathcal{D}_{\text{ref}}| \gg n$. For an unlabelled batch $x_1, \ldots, x_n$ let T_n and s_n be the mean and sample standard deviation of $u(x_i)$. The test of $H_0 : \mu_{\text{deploy}} \le \mu_0 + m$ against $H_1 : \mu_{\text{deploy}} > \mu_0 + m$ rejects when $T_n > \mu_0 + m + z_\alpha s_n/\sqrt{n}$; using s_n rather than σ_0 keeps the size correct when the deployment variance exceeds σ_0^2.

Theorem 3 (Mean-shift monitor)**.** *Let $X_1, \ldots, X_n$ be i.i.d. from P_{deploy} with $\mathbb{E}[u(X)^2] < \infty$ and standard deviation $\sigma_1(u)$, and let $\Delta(u) = \mu_{\text{deploy}}(u) - \mu_0(u) - m$. (a)* Size: *under H_0, $\lim_{n\to\infty} \Pr(\textit{reject}) \le \alpha$. (b)* Power: $\Pr(\textit{reject}) = \Phi\big(\sqrt{n}\,\Delta(u)/\sigma_1(u) - z_\alpha\big) + o(1)$. *(c)* Comparison: *if $\Delta(u_S) \ge \Delta(u_T) - \varepsilon > 0$ and $\sigma_1(u_S) \le \sigma_1(u_T)$, the student's test has asymptotic power at least $\Phi\big(\sqrt{n}\,(\Delta(u_T) - \varepsilon)/\sigma_1(u_T) - z_\alpha\big)$. (d)* Sample size: *power $1-\beta$ needs $n \gtrsim (z_\alpha + z_\beta)^2\, \sigma_1(u)^2/\Delta(u)^2$, so under the premises of (c) the student needs at most $(z_\alpha + z_\beta)^2 \sigma_1(u_T)^2/(\Delta(u_T) - \varepsilon)^2$ samples.*

Proof. (a) At the boundary $\mu_{\text{deploy}} = \mu_0 + m$, $\sqrt{n}(T_n - \mu_0 - m)/\sigma_1 \rightsquigarrow \mathcal{N}(0,1)$ by the central limit theorem and $s_n \to \sigma_1$ in probability, so by Slutsky's theorem (van der Vaart, 1998) the rejection probability tends to α; inside H_0 it is smaller. (b) $\sqrt{n}(T_n - \mu_0 - m)/s_n = \sqrt{n}(T_n - \mu_{\text{deploy}})/s_n + \sqrt{n}\,\Delta/s_n$; the first term converges to $\mathcal{N}(0,1)$ and $s_n \to \sigma_1$. (c) Power is increasing in the standardised shift Δ/σ_1, and the premises give $\Delta(u_S)/\sigma_1(u_S) \ge (\Delta(u_T) - \varepsilon)/\sigma_1(u_T)$. (d) Solve $\sqrt{n}\,\Delta/\sigma_1 - z_\alpha \ge z_\beta$ (Lehmann & Romano, 2022). □

The premises of (c) concern the deployment stream and can be checked on it (App. F.3.1).

A.6 Additional results

Error transfer.

Lemma 1 (Error transfer)**.** *For the KRR student and any comparator $u \in \mathcal{H}_k$, (a) $\|u_S - u^\star\| \le \|u_S - u_T\| + \|u_T - u^\star\|$; (b) $\|u_S - u^\star\| \le \|u - u^\star\| + 2\|u_T - u^\star\| + \sqrt{\lambda}\,\|u\|_{\mathcal{H}_k}$.*

Proof. (a) is the triangle inequality. For (b), optimality of u_S gives $\|u_S - u_T\|^2 \le \|u_S - u_T\|^2 + \lambda\|u_S\|_{\mathcal{H}_k}^2 \le \|u - u_T\|^2 + \lambda\|u\|_{\mathcal{H}_k}^2$, hence $\|u_S - u_T\| \le \|u - u_T\| + \sqrt{\lambda}\|u\|_{\mathcal{H}_k}$. Combine with (a) and $\|u - u_T\| \le \|u - u^\star\| + \|u_T - u^\star\|$. □

AUROC.

Theorem 4 (AUROC from L^2 error)**.** *If $\Pr\big(u^\star(X_1) - u^\star(X_0) > \gamma\big) \ge 1 - \alpha_0$ for independent $X_0 \sim P_0$, $X_1 \sim P_1$, then $\mathrm{AUROC}(u_S) \ge 1 - \alpha_0 - 2\big(\|u_S - u^\star\|_{L^2(P_0)}^2 + \|u_S - u^\star\|_{L^2(P_1)}^2\big)/\gamma^2$.*

Proof. Let $E_k = u_S(X_k) - u^\star(X_k)$. On the margin event, u_S ranks X_1 above X_0 unless $|E_1 - E_0| > \gamma$, and by Markov's inequality $\Pr(|E_1 - E_0| > \gamma) \le \mathbb{E}[(E_1 - E_0)^2]/\gamma^2 \le 2(\mathbb{E}E_0^2 + \mathbb{E}E_1^2)/\gamma^2$. □

The same bound holds for u_T; it is tighter for the student exactly when the student's L^2 error under P_0 and P_1 is smaller, which Theorem 2 characterises under P only.

Calibration of the score.

Proposition 1 (Calibration error from L^2 error)**.** $\mathrm{UCE}(u_S) := \mathbb{E}_P\big[\,|u_S(X) - \mathbb{E}[u^\star(X) \mid u_S(X)]|\,\big] \le \|u_S - u^\star\|$.

Proof. $u_S - \mathbb{E}[u^\star \mid u_S] = \mathbb{E}[u_S - u^\star \mid u_S]$; apply Jensen's inequality and the tower rule, then Cauchy–Schwarz. □

Binned estimators of UCE and of the expected calibration error (ECE) are biased and depend on the binning (Kumar et al., 2019). For predictive calibration we therefore also report the squared kernel calibration error (SKCE), whose unbiased estimator can take small negative values when the true value is near zero (App. B.2).

A.7 Two-Stage Training: An Information-Theoretic Account

Table 36 compares two-stage training with joint training of the prediction head and the uncertainty head on CIFAR-10. Joint training is worse on every metric: accuracy is lower by 0.92%, Pearson fidelity by 0.169 and mean OOD AUROC by 0.048. This subsection gives an informal information-theoretic account of this result; it is an argument, not a proof.

Setup. Let X be the input, $H = h_\omega(X)$ the backbone representation, f_ψ the prediction head, g_ϕ the uncertainty head. The two heads have no shared parameters; the two losses interact only through the shared backbone ω, which receives gradients from both. Teacher quantities are $p(y \mid x, \mathcal{D})$, and $\mathrm{TU}/\mathrm{AU}/\mathrm{EU}$ as in §2.

The information-theoretic decomposition. Since $H = h_\omega(X)$ is a deterministic function of X, $I(H; Z \mid X) = 0$ for any Z. Applying the chain rule for mutual information to $I(H; X, p(Y \mid X, \mathcal{D}))$ in two ways and combining yields:

$$\boxed{I(H;\ X) = I\big(H;\ p(Y \mid X, \mathcal{D})\big) + I\big(H;\ X \bigm| p(Y \mid X, \mathcal{D})\big).} \tag{20}$$

The first term is the information H carries about the teacher's predictive distribution; the second is what H retains about X beyond it.

The conflict. Because $\mathcal{D}$ and ensemble parameters are fixed, every teacher uncertainty $U \in \{\mathrm{TU}, \mathrm{AU}, \mathrm{EU}\}$ is a deterministic function of X, so by the data processing inequality

$$I(H; U \mid p(Y \mid X, \mathcal{D})) \le I(H; X \mid p(Y \mid X, \mathcal{D})). \tag{21}$$

The residual term in (20) thus bounds the information H carries about U beyond the teacher's predictive distribution. In the information-bottleneck view (Tishby et al., 2000), distillation with target $p(Y \mid X, \mathcal{D})$ alone tends to shrink this term; capacity constraints (weight decay, finite width, the implicit bias of SGD (Shwartz-Ziv & Tishby, 2017; Soudry et al., 2018; Vardi, 2023)) may add to this pressure without an explicit IB objective. The uncertainty objective instead benefits from a large residual term, since it needs $I(H; U \mid p(Y \mid X, \mathcal{D}))$ to be large. Joint training with a fixed scalar weight γ sums the two gradients on the shared backbone, where they can conflict, as in multi-task learning (Yu et al., 2020; Sener & Koltun, 2018; Standley et al., 2020); here the conflict would concern the within-class variation of the features.

For the three uncertainty types: TU is a function of $p(Y \mid X, \mathcal{D})$, so $I(H; \mathrm{TU} \mid p) = 0$ and TU raises no conflict. AU is the average of the members' entropies and $\mathrm{EU} = \mathrm{TU} - \mathrm{AU}$ measures their spread; neither is a function of the average prediction alone, so the conflict can arise for both.

Two-stage training. Stage 1 optimises $\mathcal{L}_{\mathrm{pred}}(\omega, \psi)$ alone and yields backbone parameters ω^*. Stage 2 freezes ω^* and trains g_ϕ on fixed features: a regression with no gradient conflict. Freezing limits what Stage 2 can recover to what the frozen features H^* retain. Two points suggest that this limit is not severe: (i) TU is computable directly from $\sigma(z_S)$; (ii) in practice networks do not compress perfectly, so the frozen features keep some residual information $I(H^*; X \mid p(Y \mid X, \mathcal{D}))$, and feature geometry (norms, distances, local density) can carry ensemble-spread signal that the prediction loss does not require (Tian et al., 2019).

In this account, two-stage training avoids the gradient conflict, and the information available to Stage 2 for AU and EU is bounded by $I(H^*; X \mid p(Y \mid X, \mathcal{D}))$.

B Checks of the Analysis and Additional Experiments

B.1 Positioning against prior distillation work

Table 6 lists prior methods that distil posterior or ensemble uncertainty, whether they analyse the noise in the distillation targets, and whether they are evaluated on deployment tasks.

Table 6: Distillation methods for posterior or ensemble uncertainty.

Method	What is distilled	Noise in targets analysed	Deployment tasks
Bayesian Dark Knowledge (Balan et al., 2015); Vadera & Marlin (2019)	Posterior predictive distribution (online SGLD teacher)	No	No
GPED (Vadera et al., 2020)	Posterior expectations of chosen functions, incl. entropy-based uncertainty	No	No
EnD^2 (Malinin et al., 2019); Hydra (Tran et al., 2020)	Distribution over member predictions (Dirichlet output / multiple heads)	No	No
Shen et al. (2021)	Uncertainty-aware predictive distribution	No	No
Holder & Shafique (2021); DUDES (Landgraf et al., 2023)	Per-pixel ensemble uncertainty maps	No	No
CSD (Zanger et al., 2025)	Variance of the infinite-width NTK GP of an ensemble	No	No
Ours	Scalar ensemble MI, two-stage with frozen backbone, log-space + ranking loss	Yes (§4, App. B.4)	Monitoring, routing

B.2 Ground-truth evaluation

Protocol. No model is retrained. We reuse the trained models and their cached per-sample predictions and scores, and score every method against the labels of each benchmark's standard ID test set: misclassification AUROC (each method's uncertainty ranks its own errors above its correct predictions), AURC, and accuracy at 80% coverage. Every method is scored with its own uncertainty (MI for the sampling methods and the teacher, the head output for the student, the Dirichlet epistemic term for EDL, the density or GP variance for DDU and SNGP, the predicted error for DEUP), not with $1 - \max_c p_c$. Table 2 reports the student, the teacher and the best other method, and Table 29 gives AURC for the methods run on every benchmark. Table 7 lists every method, and Table 8 gives accuracy at 80% coverage for the student and the teacher. Per-set OOD AUROC for every method is in Tables 26 and 28.

Calibration (SKCE). We estimate the squared kernel calibration error (Widmann et al., 2019) of the predictive distribution with the linear-time unbiased estimator, an RBF kernel on the probability simplex with median-heuristic bandwidth, averaged over 20 random pairings with $n = 10{,}000$ per cell. The unbiased estimator can be slightly negative when the true value is near zero. The student's values are of the same order as those of the teacher, MC Dropout, SWAG and LLLA (with its fitted temperature), and below EDL's (Table 9). The differences among the student, the teacher and MC Dropout are within the estimator's Monte Carlo noise (about $\pm 10^{-4}$). A paired bootstrap of the student–teacher difference (pairs formed by two copies of one resampled point are dropped) gives $8.7{\times}10^{-6}$ $[-5.7{\times}10^{-6}, 1.2{\times}10^{-5}]$ on MNIST, $-2.4{\times}10^{-5}$ $[-1.3{\times}10^{-4}, 7.1{\times}10^{-5}]$ on CIFAR-10 and $-2.3{\times}10^{-5}$ $[-4.2{\times}10^{-5}, 3.4{\times}10^{-5}]$ on Tiny ImageNet; every 95% interval contains zero.

B.3 Fidelity against the variability of the teacher

A retrained teacher does not reproduce the original one exactly, so the agreement between two independently trained teachers is a natural scale for fidelity. On the test inputs covered by the independent pools (first $1{,}000$ CIFAR-10 and first $2{,}000$ MNIST test images) we compare the student's correlation with its teacher to the correlation between two pools (Table 10). On CIFAR-10 the student's Spearman correlation with the teacher matches the pool-to-pool value (difference -0.012 $[-0.041, 0.018]$) and its Pearson correlation is lower (-0.079 $[-0.120, -0.037]$). On MNIST the student's correlations exceed the pool-to-pool values; the MNIST pools have only 5 members, so their mutual agreement is lower than that of 16-member teachers. On CIFAR-100, where the teacher and the pools all have 5 members (full test set), the student's correlations with its teacher (0.798 /

Table 7: Ground-truth utility of every method on the ID test sets: accuracy, AURC (lower is better) and misclassification AUROC, each with the method's own uncertainty score.

Benchmark	Method	Acc. (%)	AURC ↓	Miscls. AUROC ↑
MNIST	Teacher	99.54	0.00011	0.979
	Ours	99.51	0.00015	0.973
	MC Dropout	99.17	0.00033	0.965
	EDL	99.14	0.00022	0.980
	Laplace (LLLA)	99.20	0.00025	0.973
	SGLD	99.12	0.00028	0.973
	DDU	99.20	0.00136	0.864
	SNGP	99.44	0.00017	0.974
	DEUP	99.20	0.00021	0.978
	Energy (member)	99.20	0.00048	0.950
	ViM (student backbone)	99.51	0.00038	0.932
	ViM (member)	99.20	0.00088	0.906
	SWAG	99.31	0.00015	0.982
	SWAG ($1 - \max p$)	99.31	0.00011	0.988
CIFAR-10	Teacher	96.96	0.00235	0.947
	Ours	95.99	0.00343	0.942
	MC Dropout	93.63	0.00705	0.928
	EDL	93.51	0.00640	0.946
	Laplace (LLLA)	94.56	0.00945	0.895
	SGLD	94.66	0.01751	0.831
	DDU	94.66	0.00985	0.873
	SNGP	93.76	0.00663	0.934
	DUE	95.32	0.00856	0.892
	DEUP	94.66	0.00865	0.883
	Energy (member)	94.66	0.01703	0.863
	ViM (student backbone)	95.99	0.00477	0.915
	ViM (member)	94.66	0.00957	0.881
	SWAG	93.91	0.00649	0.933
	SWAG ($1 - \max p$)	93.91	0.00732	0.932
Tiny ImageNet	Teacher	88.13	0.02241	0.892
	Ours	87.50	0.02427	0.888
	MC Dropout	85.60	0.02972	0.882
	Laplace (LLLA)	86.51	0.02821	0.883
	EDL	84.70	0.03150	0.895
	SGLD	73.50	0.08861	0.831
	DDU	87.01	0.03855	0.825
	SWAG	87.33	0.02044	0.912
	DEUP	87.01	0.03605	0.824
	Energy (member)	87.01	0.04501	0.812
	ViM (student backbone)	87.51	0.03515	0.834
	ViM (member)	87.01	0.04289	0.810

Table 8: Accuracy at 80% coverage when abstaining on the most uncertain inputs.

	MNIST	CIFAR-10	Tiny ImageNet
Ours	100%	99.7%	95.2%
Teacher	100%	99.8%	95.7%

0.835) exceed the pool-to-pool values (0.777 / 0.825; differences $+0.021$ [0.013, 0.028] and $+0.010$ [0.004, 0.016]).

B.4 Checks of the analysis

Controlled simulation (surrogate level). Inputs are 256 equally spaced training points and $2{,}048$ test points on $[0, 1]$ with an RBF kernel (length scale 0.09). The ground truth is a positive combination of the first five empirical eigenfunctions. The teacher's error is a random combination of the first 32 eigenfunctions with coefficients $\mathcal{N}(0, (1.6\, w_i)^2/K)$, so it is correlated across inputs and its variance scales as $1/K$; the eigenvalues of these 32 modes are at least 1.1×10^{-10}, while later ones are at machine precision and their eigenfunctions vanish on the test points. The weights w ($\|w\| = 1$) decay ($e^{-i/7}$), are flat, or rise ($e^{-(31-i)/7}$) in i. The student is the spectral filter $\mu_i/(\mu_i + n\lambda)$ applied to the teacher's training values, either directly (L2) or to $\log(1 + u_T)$ followed by $\exp(\cdot) - 1$ (log-space), and is evaluated exactly. The grid covers 3 noise spectra (leading: concentrated in the leading eigenfunctions, where the signal lies; flat; trailing: concentrated in the trailing ones,

Table 9: Squared SKCE of the predictive distribution on ID test sets (closer to 0 is better).

	CIFAR-10	MNIST	Tiny ImageNet
Ours	7.7×10^{-5}	4.7×10^{-5}	-3.2×10^{-5}
Teacher	1.0×10^{-4}	3.8×10^{-5}	-8.9×10^{-6}
MC Dropout	2.0×10^{-5}	1.3×10^{-5}	-2.7×10^{-5}
EDL	4.5×10^{-4}	2.9×10^{-4}	9.7×10^{-5}
Laplace (LLLA)	-1.8×10^{-5}	1.3×10^{-5}	3.4×10^{-5}
SWAG	9.4×10^{-5}	-1.4×10^{-6}	-2.1×10^{-5}

Table 10: Pearson / Spearman correlation between uncertainty scores on test inputs, with 95% bootstrap intervals. Pool-to-pool values are averaged over pairs of independently trained pools (16 members on CIFAR-10, 5 on MNIST); the main teacher has 16 members on both.

Comparison	CIFAR-10	MNIST
Student vs. its teacher	0.800 [0.758, 0.839] / 0.790 [0.762, 0.816]	0.821 [0.799, 0.844] / 0.833 [0.817, 0.848]
Pool vs. pool	0.879 [0.851, 0.902] / 0.802 [0.775, 0.826]	0.769 [0.747, 0.795] / 0.718 [0.699, 0.736]
Main teacher vs. pool	0.903 / 0.779	0.830 / 0.775
Student vs. pool	0.792 / 0.786	0.741 / 0.738

which the filter shrinks), $K \in \{5, 16, 64\}$, 17 values of λ and 2 losses, giving 306 settings with 100 repetitions each and exact bias–variance decompositions; the best λ is located on a finer grid of 281 values. Findings: (i) the teacher's error falls as $1/K$ (1.92×10^{-3}, 6.27×10^{-4}, 1.51×10^{-4} at $K = 5, 16, 64$, flat spectrum); (ii) the best λ falls with it (7.1×10^{-4}, 2.8×10^{-4}, 7.9×10^{-5}, flat spectrum, L2; Remark 1); (iii) every fixed $\lambda \geq 1.4\times10^{-2}$ loses at every K, for every spectrum and loss; (iv) at the best λ the ratio of the student's to the teacher's error is 0.98–1.00 for the leading spectrum, 0.42–0.52 for the flat spectrum and 0.010–0.022 for the trailing spectrum, as consequences (i) and (iii) of Theorem 2 predict: noise in the directions the filter keeps passes to the student, and noise in the directions it shrinks is removed; (v) over the λ grid the student is strictly better in 180 of 306 settings, and the two losses behave alike in all three spectra. With flat noise at $K = 5$ and the best λ (7.1×10^{-4}) the teacher's squared bias is 1.6×10^{-5} and its variance 1.90×10^{-3}; the student's error is 8.1×10^{-4}, lower than the teacher's on 100% of test points.

Structure of the teacher's error (direct). *Nested sub-ensembles.* Taking $K = 3, 5, 10$ as prefixes of one 16-member CIFAR-10 pool, the error fields have effective rank 2.34 (independent errors would give 3, identical errors 1) and pairwise correlations 0.38–0.57. These prefixes share members, so the correlation is partly mechanical. *Independent pools.* We therefore trained five independent 5-member pools on MNIST and three independent 16-member ResNet-18 pools on CIFAR-10 from scratch with disjoint seeds, and measured each pool's MI on the first 2,000 (MNIST) or 1,000 (CIFAR-10) test images against a reference built from the other pools. With a reference made of the other pools' first K members, the error follows $\mathrm{MSE}(K) = 5.0\times10^{-5} + 6.4\times10^{-4}/K$ on MNIST ($R^2 = 0.94$) and $1.83\times10^{-3} + 2.18\times10^{-2}/K$ on CIFAR-10 ($R^2 = 0.93$). Cross-pool error correlations are -0.14 (MNIST) and -0.47 (CIFAR-10) on average, so the positive correlation among prefixes came from shared members. Because that reference also grows with K, we repeated the fit with a fixed reference of all other members (20 on MNIST, 32 on CIFAR-10): R^2 rises to 0.98–0.99 and the CIFAR-10 intercept is 1.68×10^{-3}. Splitting the reference in two halves estimates its own noise (7.6×10^{-4}); half the variance of the difference between two full pools, minus the member-level term b'/M (b' the fitted slope, M=16), estimates a pool-level component $C_0 = 4.9\times10^{-4}$; the remaining 4.3×10^{-4} is not attributed and includes the reference's own pool-level deviation (about $C_0/2$). On MNIST the intercept (4.9×10^{-5}) is almost all reference noise (3.0×10^{-5}) and the estimate of C_0 is negative, i.e. not detectable. Remark 2 explains why the bias β_K itself cannot be identified this way.

Finite-K comparison with trained heads. On CIFAR-10 we use one independent pool's first K members as the teacher, $K \in \{2, 3, 5, 10, 16\}$, train Stage-2 heads on its MI over the auxiliary set (log-space loss with and without the ranking term, weight decay λ on a grid; and ridge regression on the frozen features), select λ per K, and measure the error of teacher and student on the test inputs, on the $\log(1 + u)$ scale, against the fixed 32-member reference. Table 11 gives the result. In squared error, the student is better at $K = 2$ for all three students, at $K = 3$ for two of them, about equal at

$K = 5$, and worse from $K = 10$ on, and the selected regularisation does not grow with K. Both agree with Theorem 2 and Remark 1 for a teacher whose noise shrinks with K while the student's bias does not. Rank correlation behaves differently (Table 12): on the ID test inputs the student ranks closer to the reference than its teacher at every K, while within the OOD sets it does so only for small K. At $K = 16$ the student's ID ranking matches the teacher's, so its larger squared error reflects its values rather than its ordering.

Table 11: Teacher minus student error ($\times 10^{-3}$, $\log(1 + u)$ scale, test inputs) against a 32-member reference on CIFAR-10, at the best λ for each K (in parentheses). Positive: the student is closer to the reference than its teacher.

Student	K=2	K=3	K=5	K=10	K=16
Head, log + ranking	$+1.27$ (10^{-5})	$+0.54$ (10^{-4})	-0.13 (10^{-5})	-0.90 (0)	-1.46 (0)
Head, log only	$+1.42$ (10^{-4})	-0.07 (10^{-4})	-0.28 (10^{-5})	-0.91 (10^{-5})	-1.35 (10^{-5})
Ridge	$+1.69$ (1)	$+0.45$ (1)	-0.01 (1)	-0.87 (0.3)	-1.42 (0.1)

Table 12: Spearman correlation with the 32-member reference (teacher / student, head with log-space and ranking loss, no weight decay), per split.

Split	K=2	K=3	K=5	K=10	K=16
CIFAR-10 test	0.59 / 0.71	0.65 / 0.79	0.72 / 0.81	0.78 / 0.83	0.83 / 0.84
SVHN	0.31 / 0.42	0.32 / 0.45	0.44 / 0.50	0.59 / 0.53	0.65 / 0.55
CIFAR-100	0.51 / 0.60	0.56 / 0.57	0.65 / 0.62	0.75 / 0.63	0.80 / 0.64

All three from-scratch benchmarks, with intervals. We repeated the comparison without weight decay on MNIST (teacher: one 5-member pool; reference: the other 20 members) and on CIFAR-100 (teacher: one 5-member pool; reference: the other two pools, 10 members), with both losses, and computed paired-bootstrap intervals over test inputs (Table 13). The ratio of the student's to the teacher's squared error is higher at the largest K than at $K = 2$ for every benchmark and loss. With the training loss of §3.3, the student's Spearman correlation with the reference exceeds the teacher's at every K on CIFAR-10 and MNIST and is lower for $K \geq 3$ on CIFAR-100. With either of the other two CIFAR-100 pools as the teacher (log + rank) the ratios at $K = 2, 3, 5$ are 1.11, 1.20, 1.42 and 1.30, 1.53, 1.62.

Table 13: Student / teacher squared error against an independent reference, and the student-minus-teacher Spearman correlation with it (test inputs, $\log(1 + u)$ scale, no weight decay; 95% paired-bootstrap intervals). "Log + rank" is the training loss of §3.3; "log" drops the ranking term. On MNIST these heads use a Softplus output with its bias initialised at the inverse Softplus of the mean target, Adam with learning rate 10^{-3} and the losses listed, not the plain MSE of the reported MNIST head (§3.2). With the ranking term their outputs spread more than the teacher's MI (standard deviation 0.028 vs. 0.015 at K=2), which raises the squared-error ratio while the Spearman correlation improves.

Benchmark	Loss	K=2	K=3	K=5	K=10	K=16
MSE ratio (student / teacher)						
CIFAR-10	log + rank	0.83 [0.80, 0.86]	0.93 [0.89, 0.96]	1.04 [0.99, 1.09]	1.36 [1.28, 1.44]	1.81 [1.68, 1.93]
CIFAR-10	log	0.88 [0.85, 0.91]	1.09 [1.05, 1.13]	1.18 [1.13, 1.25]	1.49 [1.40, 1.59]	1.88 [1.76, 2.02]
MNIST	log + rank	1.75 [1.60, 1.90]	2.71 [2.45, 2.97]	8.43 [7.58, 9.36]	–	–
MNIST	log	0.86 [0.81, 0.91]	0.96 [0.89, 1.05]	1.18 [1.05, 1.31]	–	–
CIFAR-100	log + rank	0.99 [0.97, 1.02]	0.98 [0.94, 1.02]	1.06 [1.02, 1.10]	–	–
CIFAR-100	log	1.33 [1.29, 1.37]	1.55 [1.49, 1.61]	1.75 [1.67, 1.82]	–	–
Spearman correlation with the reference, student minus teacher						
CIFAR-10	log + rank	+0.122 [0.114, 0.131]	+0.138 [0.130, 0.147]	+0.088 [0.080, 0.095]	+0.046 [0.039, 0.053]	+0.011 [0.005, 0.018]
CIFAR-10	log	+0.118 [0.109, 0.129]	+0.028 [0.019, 0.038]	−0.025 [−0.035, −0.017]	−0.071 [−0.081, −0.062]	−0.108 [−0.118, −0.099]
MNIST	log + rank	+0.077 [0.064, 0.091]	+0.082 [0.071, 0.095]	+0.023 [0.014, 0.032]	–	–
MNIST	log	−0.039 [−0.055, −0.024]	−0.034 [−0.048, −0.019]	−0.096 [−0.107, −0.084]	–	–
CIFAR-100	log + rank	−0.006 [−0.014, 0.003]	−0.020 [−0.027, −0.013]	−0.029 [−0.034, −0.023]	–	–
CIFAR-100	log	−0.027 [−0.036, −0.019]	−0.036 [−0.042, −0.029]	−0.044 [−0.049, −0.038]	–	–

Condition of Theorem 2. Because the bias β_K cannot be identified from pools (Remark 2), we check the unbiased form of the condition, $\mathrm{tr}[(I - R_\lambda^2)C_K] > \|(I - R_\lambda)u_{\mathrm{ref}}\|^2$, with C_K estimated from the pools and the pooled reference in place of $u^\star$, in the eigenbasis of the trained head's empirical NTK on CIFAR-10. C_K is represented by its diagonal in that basis, estimated from the three pools' errors at $K = 16$ (so its rank is at most three), the reference uses all 48 members, and λ is on a

grid of 33 values relative to the largest eigenvalue. It holds for 13 of 33 values of λ: throughout $\lambda/\lambda_{\max} \in [10^{-6}, 10^{-3}]$, and not above about 1.8×10^{-3}.

Moment audit of trained heads (direct). For each benchmark's ID test set and each OOD set we compute the exact means, variances and second moments of the teacher's and the student's EU from cached outputs. This is an analogue of Theorem 1 on the EU scale and on test distributions rather than P, and it is what premise (ii) of Corollary 1 needs. Intervals are paired bootstrap percentiles (1,000 resamples). On the $v = \log(1 + u)$ scale of the analysis the count of contracting splits is the same. On the 21 vision splits the variance contracts on 17; for example CIFAR-10 ID $0.0127 \to 0.0097$ (ratio 0.76) and CIFAR-10→DTD $0.0340 \to 0.0121$ (ratio 0.36). On MNIST and CIFAR-10 the only exception is MNIST ID, where the teacher's variance is small (4.2×10^{-4}) and the student's second moment also expands; there the trained head departs from the surrogate. The other three exceptions are Tiny ImageNet ID, →MNIST and →STL-10. The mean term of Eq. (11) exceeds 20% of the second-moment difference on 20 of the 21 splits, so it cannot be neglected. On the six matched SST-2 splits (ID and five OOD sets; moment-identity residuals at most 1.2×10^{-9}) the student's variance exceeds the teacher's by factors of 3.5–5.5, with all bootstrap intervals excluding zero. On CIFAR-100 (K=5, OpenOOD sets) the variance contracts on all 7 splits on both scales, with every interval excluding zero. The overall count is 24 of 34. All ten non-contracting splits have small teacher variance (SST-2: 2.3–4.7×10^{-4}). A BERT-base student, the teacher's own backbone with LoRA and otherwise the same recipe, has variance ratios of 4.6–10.7 on the same splits, and the DistilBERT student trained with the same protocol (validation split from the training set) has 4.7–10.1, so the expansion persists with the teacher's backbone. The SST-2 teacher is fine-tuned, and these splits are not used to test the analysis (§5.1).

Shrinkage per component (direct). Coordinate-wise shrinkage $\rho_j \in (0, 1)$ holds only in the kernel's own eigenbasis. In any other orthonormal basis Q the filter is not diagonal, and a contraction ($\|R_\lambda u\| \le \|u\|$) can enlarge individual coordinates $|(Q^\top R_\lambda u)_j| > |(Q^\top u)_j|$. Per-component ratios in a PCA basis of the head's input features (Figure 5) are therefore only a heuristic diagnostic. We computed the empirical NTK of the trained Stage-2 head on 1,000 CIFAR-10 test inputs and projected both scores on its eigenvectors. 52.3% of components have student/teacher energy ratio above 1 (PCA basis: 25.6%), so the trained head is not a coordinate-wise shrinker even in its own kernel basis. In aggregate it contracts: weighted by teacher energy the ratio is 0.745 (total energy $13.64 \to 10.16$). The largest ratios (99th percentile 667) occur where the teacher's energy is near zero and the ratio is ill-conditioned; the two highest-energy components (59% of the energy) have ratios 1.17 and 0.59.

Premises of the downstream results. Counting ID/OOD pairs, the variance premise of Corollary 1 holds for all 6 CIFAR-10 and all 6 CIFAR-100 pairs and for none of the 5 MNIST pairs, because the MNIST ID split does not contract; it also fails for the 7 Tiny ImageNet and 5 SST-2 pairs, and these benchmarks are not used to test the analysis. So premise (ii) holds on CIFAR-10 and CIFAR-100. For premise (i), the student's ID/OOD mean gap is smaller than the teacher's ($\varepsilon > 0$) on 6 of 6 CIFAR-10 and 5 of 6 CIFAR-100 pairs. With the measured ε, the student's standardised separation $\Gamma(u_S)$ is above the bound of Eq. (17) on all 12 pairs, and above the teacher's $\Gamma(u_T)$ on 9 of 12 (EU scale; on the $\log(1 + u)$ scale the bound also holds on all 12 pairs). For monitoring, the premises of Theorem 3(c) are checked per scenario in App. F.3.1.

B.5 Teacher member diversity on Tiny ImageNet

The main Tiny ImageNet teacher draws, per member, the LoRA rank ($\{4, 8, 16, 32\}$), target modules, number of unfrozen blocks, augmentation, label smoothing, data subsample and optimiser settings. We trained a second 16-member teacher whose members share one recipe (rank 8, attention projections, no unfrozen blocks, fixed augmentation) and differ only by seed, distilled it into its own student, and compared both pairs on the same seven OOD sets. The comparison is confounded: the shared-recipe members were trained for 80 instead of 50 epochs, its student used a different Stage-1 recipe, and the two teachers were trained in different runs.

With diverse members, symmetric member KL is about 3× larger and the ID student–teacher Pearson correlation is higher by 0.076. The differences are not uniform across sets: the CIFAR-10 shift is slightly worse for both teachers and the FashionMNIST student slightly worse, while the largest difference is the student on SVHN (+0.118).

Table 14: Effect of member diversity on Tiny ImageNet.

	Shared recipe	Main teacher (diverse members)
Member disagreement (relative)	1.00	1.62
Teacher mean OOD AUROC	0.883	0.897
Student mean OOD AUROC	0.834	0.860

B.6 TIMED COST

One full CIFAR-10 ($K{=}16$) pipeline was timed on an RTX 4090 in a separate run with the reported schedule (200 Stage-1 and 100 Stage-2 epochs; Table 15). This run gives Stage-1 accuracy 95.93% and Stage-2 Spearman 0.800 (reported models: 95.99%, 0.799). With offline cost C (seconds) and deployment throughputs r_T and r_S (samples/s), the student pays for itself after $N^\star = C/(1/r_T - 1/r_S)$ samples: with the throughputs of Table 33, $N^\star = 1875.5/(1/2073 - 1/33088) \approx 4.15$M, i.e. about 33 minutes of traffic at the ensemble's rate.

Table 15: Offline cost of the CIFAR-10 ($K{=}16$) pipeline.

Component	Wall time	GPU-hours
Teacher targets (train + auxiliary inputs)	185.8 s	0.052
Stage 1 distillation (200 epochs)	1,258.4 s	0.350
Stage 2 head (100 epochs)	431.3 s	0.120
Total	1,875.5 s	0.521

B.7 HETEROSCEDASTIC REGRESSION

We train 5-member MLP ensembles (two hidden layers of width 128, mean output only, members differing by seed) on a 1-D toy task with heteroscedastic noise and a gap in the training inputs, on Concrete and on Energy (random $75/25$ splits, one seed), and distil the variance of the members' mean predictions, i.e. the ensemble's epistemic variance, with the same two stages (a Softplus head on the frozen student; log-space regression plus ranking with weight 0.25). Selective risk abstains in decreasing order of each score.

Table 16: Distilling the predictive variance of regression ensembles.

Dataset	Pearson ↑	Spearman ↑	Selective risk ↓ (student / teacher score)
Toy	0.665	0.784	0.122 / 0.106
Concrete	0.664	0.527	0.094 / 0.096
Energy	0.711	0.839	0.0026 / 0.0026

The student's outputs are non-negative, and its mean score is above the teacher's (0.032 vs. 0.0006 on Energy): at such small targets the Softplus output is biased upward under the log-space loss. The selective-risk results are mixed.

B.8 MODEL SELECTION AND A HELD-OUT PROTOCOL

Checkpoint selection. Each training run keeps one checkpoint (epoch); no model is chosen among seeds. The criterion is prediction accuracy or label-free agreement with the teacher, never an uncertainty metric (AURC, misclassification AUROC or OOD AUROC). Some of these choices use the evaluation data. The MNIST and CIFAR-10 teacher members and the CIFAR-10 Stage-1 student keep the epoch with the highest test accuracy. On VOC and SST-2 the validation or dev set is also the evaluation set; there the VOC teacher members and the Stage-1 students of both benchmarks keep the epoch with the highest mIoU or accuracy on that set. On every benchmark, Stage 2 keeps the epoch with the highest Spearman correlation between the head and the teacher's MI on the evaluation inputs; no labels are used, but evaluation inputs are.

Held-out protocol. Three choices in the reported models may favour them. (i) The Stage-2 checkpoint is selected on the test inputs, as described above. (ii) On Tiny ImageNet the Tier-2 inputs are corrupted copies of the first images of the validation split, which also serves as the ID test set. (iii) On CIFAR-10 the Tier-3 mixup images are clipped to $[0, 1]$ in normalised space, which sets about half of their pixel values to zero. We also train Stage 2 on the same frozen backbones with a held-out protocol: 5,000 training images (CIFAR-10, MNIST) or 10 per class (Tiny ImageNet) are held out for checkpoint selection and removed from every tier, the auxiliary inputs are stored and their targets computed on exactly those tensors. Table 17 compares the results.

Table 17: Reported models versus Stage 2 retrained with the held-out protocol (single seed; five-seed spread in the text).

Benchmark	Run	Mean OOD AUROC ↑	AURC ↓	Pearson / Spearman
CIFAR-10	Reported	0.9010	0.00343	0.815 / 0.799
	Held-out, unclipped mixup	0.9036	0.00335	0.815 / 0.802
	Held-out, clipped mixup	0.9024	0.00343	0.813 / 0.794
Tiny ImageNet	Reported	0.8604	0.02427	0.842 / 0.885
	Held-out, Tier 2 from training images	0.8558	0.02459	0.814 / 0.871
MNIST	Reported (linear head, MSE)	0.9913	0.00015	0.818 / 0.805
	Softplus head, log-space loss, output bias initialised	0.9923	0.00014	0.751 / 0.797

On CIFAR-10 and Tiny ImageNet the ground-truth metrics change by at most 0.005; on Tiny ImageNet the correlation with the teacher drops by 0.028. Over five Stage-2 seeds under the held-out protocol, mean OOD AUROC is 0.9041 ± 0.0006 (CIFAR-10) and 0.8584 ± 0.0038 (Tiny ImageNet), and Pearson correlation 0.816 ± 0.001 and 0.812 ± 0.003. Retraining both stages on CIFAR-10 with the released code and the held-out protocol (three seeds) gives accuracy $95.75 \pm 0.09\%$, AURC 0.0037 ± 0.0001, mean OOD AUROC 0.905 ± 0.004 and Pearson 0.813 ± 0.001; the Stage-1 seed changes mean OOD AUROC more than the Stage-2 seed does. The held-out images are removed from the student's training data only: the teacher, the same as in the main results, was trained on all training images.

On MNIST the reported head has a linear output, clamped at zero at test time, and is trained with MSE. With the recipe of the other benchmarks (Softplus output, log-space loss) and a zero-initialised output bias, the Softplus saturates and the head outputs a constant in four of five seeds. Initialising the output bias at the inverse Softplus of the mean target avoids this in four of five seeds and matches the reported model's ground-truth metrics (0.9923 ± 0.0005 mean OOD AUROC over the four). Spearman correlations here use average ranks for ties; the reported MNIST head outputs exact zeros on 53% of test inputs, and breaking ties arbitrarily gives 0.748 instead of 0.805.

B.9 Distillation baselines

EnD^2. Following Malinin et al. (2019), the network outputs Dirichlet concentrations $\alpha(x) = \exp(z(x))$ and is trained by the negative log-likelihood of the 16 members' predictive distributions under the Dirichlet, with central smoothing ($\epsilon = 10^{-4}$) and temperature annealing from 10 to 1 over the first half of training, applied to both member logits and concentrations as in the official code; log-concentrations are clipped to $[-10, 10]$. The uncertainty score is the Dirichlet's mutual information and the prediction is α/α_0. The full version trains a ResNet-18 from scratch with the Stage-1 optimiser on the same clean and auxiliary inputs as the student; the head version trains a head of the same size as the uncertainty head on the student's frozen Stage-1 backbone, so that only the output parametrisation and loss differ.

Hydra. Following Tran et al. (2020), a shared body carries 16 heads, each matched to one member by temperature-scaled KL after a first stage that trains a single head on the ensemble mean and copies it. The light version shares the whole ResNet-18 and uses $512 \to 256 \to 10$ MLP heads; the block version shares the stem and the first three stages and gives each head one residual block and a linear layer, as in the original. Hydra has no official code; each stage runs 200 epochs with the Stage-1 optimiser on clean training images, the original recipe. The uncertainty score is the MI of the 16 heads. Hydra is the most accurate student in Table 3, but its MI is a weaker error and OOD score than the student's, and its fidelity to the teacher's MI is lower.

B.10 OTHER BAYESIAN TEACHERS

The recipe needs only the teacher's uncertainty on the training and auxiliary inputs. On CIFAR-10 we replaced the ensemble with three other teachers: MC Dropout (p=0.2 before the final layer, trained from scratch for 200 epochs, BatchNorm in evaluation mode, 16 passes), SWAG (20 weight samples, BatchNorm statistics recomputed per sample) and last-layer Laplace (LLLA, 20 samples). For each, Stage 2 is trained on that teacher's MI on the frozen student backbone, whose Stage 1 is distilled from the ensemble (three seeds). For MC Dropout we also retrained both stages from the dropout teacher alone (one seed). Table 18 reports the results. The student's mean OOD AUROC exceeds each non-ensemble teacher's, while its fidelity to these teachers is lower than to the ensemble.

Table 18: Students distilled from different Bayesian teachers on CIFAR-10. Student rows: mean of three Stage-2 seeds ("both stages": one seed), selected on a held-out validation split. Fidelity: Pearson / Spearman correlation with that teacher's MI. Brackets: 95% paired-bootstrap interval of student minus teacher. The MC Dropout teacher (dropout p=0.2 before the classifier, trained from scratch for 200 epochs, BatchNorm in evaluation mode) is not the MC Dropout baseline of Table 7 (Dropout2d p=0.1 after residual stages 1–3, fine-tuned for 20 epochs from one member, BatchNorm on test-batch statistics). The LLLA teacher is the same fit as the LLLA baseline, with 20 instead of 16 samples and no temperature.

Teacher	Model	Acc.	Fidelity	AURC ↓	Miscls. AUROC	Mean OOD AUROC
Ensemble (K=16)	Teacher	96.96%	–	0.00235	0.947	0.903
	Ours	95.99%	0.815 / 0.801	0.00335	0.943	0.904 [−0.001, 0.003]
MC Dropout (16 passes)	Teacher	95.55%	–	0.00531	0.927	0.873
	Ours	95.99%	0.488 / 0.581	0.00515	0.911	0.897 [0.021, 0.027]
	Ours, both stages	95.36%	0.501 / 0.597	0.00615	0.914	0.879 [0.003, 0.009]
SWAG (20 samples)	Teacher	93.91%	–	0.00649	0.933	0.863
	Ours	95.99%	0.425 / 0.678	0.00406	0.931	0.915 [0.048, 0.056]
Laplace (LLLA, 20 samples)	Teacher	94.65%	–	0.00912	0.897	0.866
	Ours	95.99%	0.479 / 0.559	0.00623	0.892	0.887 [0.018, 0.024]

B.11 CIFAR-100

The teacher consists of five ResNet-18 members trained from scratch for 200 epochs on 45,000 training images with the CIFAR-10 recipe and per-member random hyperparameters (test accuracy 72.1–76.3%); the remaining 5,000 images are the validation split for all models. The student follows the CIFAR-10 recipe with a 612-dimensional head input. OOD sets follow OpenOOD v1.5 (near: CIFAR-10, Tiny ImageNet; far: MNIST, SVHN, Textures, Places365). Table 19 lists all methods. The student exceeds its teacher on mean OOD AUROC (+0.011 [0.007, 0.015]) and ties it on misclassification AUROC, with higher AURC (+0.005 [0.003, 0.008]) from its lower accuracy. Energy on the student's backbone has the highest mean OOD AUROC; the EnD^2 head has the lowest AURC and the highest fidelity. The student's variance is below the teacher's on all seven splits. With two further 5-member pools, the fixed-reference fit does not separate the pool-level component from the reference noise.

Table 19: CIFAR-100 (K=5). Student: mean over three Stage-2 seeds (standard deviation of mean OOD AUROC 0.004). Post-hoc scores on one teacher member's backbone (member) or on the student's.

Method	Acc.	AURC ↓	Miscls. AUROC	Fidelity	Near-OOD	Far-OOD	Mean OOD
Teacher (K=5)	80.78%	0.0539	0.846	–	0.787	0.800	0.796
Ours	78.99%	0.0593	0.853	0.798 / 0.835	0.793	0.814	0.807
MC Dropout (16 passes)	77.55%	0.0667	0.851	0.702 / 0.765	0.780	0.775	0.777
DDU (member)	72.14%	0.2021	0.587	0.175 / 0.208	0.564	0.653	0.623
Energy (member)	72.14%	0.0947	0.846	0.722 / 0.820	0.769	0.719	0.736
Energy (student backbone)	78.99%	0.0639	0.847	0.757 / 0.781	0.814	0.812	0.813
ViM (member)	72.14%	0.1695	0.652	0.383 / 0.426	0.594	0.639	0.624
ViM (student backbone)	78.99%	0.0929	0.749	0.626 / 0.658	0.732	0.782	0.765
EnD^2 head (student backbone)	79.01%	0.0538	0.875	0.810 / 0.855	0.807	0.793	0.798

B.12 OpenOOD v1.5

We evaluated the trained CIFAR-10 models, without retraining, on the OpenOOD v1.5 benchmark (Zhang et al., 2024) with its data lists, preprocessing and metric code (near-OOD: CIFAR-100, Tiny ImageNet; far-OOD: MNIST, SVHN, Textures, Places365). OpenOOD's own Energy and ViM postprocessors reproduce our numbers for one member's backbone.

Table 20: OpenOOD v1.5, CIFAR-10: mean AUROC over near- and far-OOD sets.

Method	Near-OOD ↑	Far-OOD ↑
Teacher (K=16)	0.930	0.954
Ours	0.930	0.951
ViM, student backbone	0.917	0.952
Energy, student backbone	0.915	0.938
EDL	0.909	0.925
MC Dropout (16 passes, BN statistics from the test batch, shuffled batches)	0.898	0.938
MC Dropout (16 passes, BN in evaluation mode)	0.904	0.917
SNGP	0.898	0.925
ViM, one member's backbone	0.869	0.929
DDU	0.858	0.919
Energy, one member's backbone	0.832	0.871

MC Dropout is reported in two variants: with BatchNorm statistics from each shuffled test batch, as in the other tables, and with BatchNorm in evaluation mode.

C Experimental Protocol and Baseline Scope

This section gives the hyperparameters (§C.1), a summary of the four distance-aware single-pass baselines (§C.2), and per-baseline win/tie/loss counts on the three primary metrics (§C.3). The baseline settings are listed in §D.8.

C.1 Hyperparameters

Teachers. MI is computed in nats from the members' softmax outputs (Eq. 3); the teacher's prediction and the Stage-1 target use the averaged distribution $\bar{p}$. Members differ in seed (initialisation and data order) and in hyperparameters drawn at random per member. MNIST (K=16, 4-layer CNN, Adam, 50 epochs): augmentation subset, label smoothing, dropout, data subsample (80–100%), learning rate, weight decay, schedule, final-layer initialisation scale. CIFAR-10 (K=16, ResNet-18, SGD, 200 epochs): learning rate, weight decay, augmentation subset, data subsample (drawn from 80–100%), initialisation scale, label smoothing, dropout, schedule. On MNIST and CIFAR-10 each member keeps the epoch with the highest test accuracy (§5.1). Tiny ImageNet (K=16, DeiT-S, 50 epochs): LoRA rank $\{4, 8, 16, 32\}$ and α, target modules (attention only, attention and projection, or also MLP), unfrozen blocks $\{0, 1, 2, 4\}$, augmentation, label smoothing, data subsample, optimiser settings. VOC (K=5, SegFormer-B2, LoRA rank 16) and SST-2 (K=5, BERT-base, LoRA rank 8): initialisation scale, label smoothing, augmentation or attention dropout, weight decay or learning rate.

Evaluation sets. OOD sets are FashionMNIST (Xiao et al., 2017), Omniglot (Lake et al., 2015), EMNIST-Letters, CIFAR-10 and SVHN (Netzer et al., 2011) for MNIST; SVHN, CIFAR-100 (Krizhevsky & Hinton, 2009), MNIST, FashionMNIST, STL-10 (Coates et al., 2011) and DTD (Cimpoi et al., 2014) for CIFAR-10; and SVHN, CIFAR-10/100, STL-10, DTD, FashionMNIST and MNIST for Tiny ImageNet. CIFAR-100 uses the OpenOOD v1.5 sets: CIFAR-10 and Tiny ImageNet (near) and MNIST, SVHN, Textures and Places365 (far).

Stage 1 and Stage 2. Table 21 lists the settings of the reported models. Stage 1 minimises Eq. (4) with target $\sigma(\log \bar{p}/\tau)$; on VOC the target is the logits of one member (member 0), computed online. The warm-up lasts 10 (CIFAR-10) or 5 (MNIST) epochs, rising linearly from 10^{-3} times the learning rate; Stage 1 uses random crops and horizontal flips on CIFAR-10 and rotations of up to $\pm 10^{\circ}$ on MNIST. The MNIST head uses LeakyReLU with slope 0.1.

Table 21: Training settings of the reported students. Stage 2 always uses β=1.

	Stage 1 (predictive distillation)	Stage 2 (uncertainty head, backbone frozen)
MNIST	Adam, lr 10^{-3}, wd 10^{-4}, warm-up + cosine, 50 ep, batch 128, α=0.7, τ=4	Adam, lr 5×10^{-3}, cosine, 100 ep, batch 128
CIFAR-10	SGD (momentum 0.9), lr 0.1, wd 5×10^{-4}, warm-up + cosine, 200 ep, batch 128, α=0.7, τ=4	Adam, lr 10^{-3}, cosine, 100 ep, batch 128
Tiny ImageNet	AdamW, lr 10^{-4}, wd 0.1, 50 ep, batch 32, 12 blocks unfrozen (backbone lr $\times$0.01, layer decay 0.75), label smoothing 0.1, MixUp/CutMix, EMA, α=0.7, τ=2	Adam, lr 3×10^{-3}, cosine, 80 ep, batch 256
VOC	AdamW, lr 6×10^{-5}, wd 0.01, poly schedule, best of 22.9k steps, batch 8, α=0.7, τ=4	Adam, lr 10^{-3}, cosine, 73 ep, batch 8
SST-2	AdamW with layer decay 0.85, lr 2×10^{-5}, wd 0.05, early stopping (best epoch 9), batch 32, α=0.5, τ=2	Adam, lr 2×10^{-4}, cosine, up to 200 ep, batch 32

Heads. One hidden layer in every case. MNIST: input $[\mathbf{h}\|\sigma(z_S)]$ (74-d), 128 hidden units, LeakyReLU, linear output clamped at zero at test time. CIFAR-10: 522-d input, 128 ReLU units, Softplus output. Tiny ImageNet: 586-d input, i.e. the 384-d class token, the 200-d softmax, and two feature-geometry inputs: the largest cosine similarity to and the smallest Euclidean distance from the class means of the frozen Stage-1 features over 30,000 clean training images, each standardised with its mean and standard deviation over clean training inputs; 256 ReLU units, Softplus output. Without these two inputs (same frozen backbone, held-out protocol of App. B.8, three Stage-2 seeds), mean OOD AUROC falls from 0.857 to 0.844, AURC rises from 0.0245 to 0.0268, misclassification AUROC falls from 0.886 to 0.873, and Pearson / Spearman fidelity fall from 0.812 / 0.871 to 0.769 / 0.849; every difference is beyond noise. VOC: 1×1 convolutions on the last encoder features (512$\rightarrow$128, BatchNorm, ReLU, 128$\rightarrow$1), Softplus, bilinear upsampling; no softmax input. SST-2: 770-d input, 256 ReLU units, dropout 0.2, Softplus output.

Losses. The ranking term (Eq. 5) draws 256 index pairs with replacement from each batch and keeps the ordered pairs with $u_T(x_i) > u_T(x_j)$ (ties dropped), with margin $\delta = 0.05$; VOC draws 512 pixel pairs per batch. On Tiny ImageNet the ranking term uses margin 0 and keeps tied pairs in the denominator, and the log-space term weights under-predictions ($g_\phi < u_T$) by 3. On MNIST the regression term is plain MSE, because the linear output can be negative.

Auxiliary inputs. All tiers are generated offline; Tier-2 corruptions use one fixed severity each (except the Tiny ImageNet blur), split equally among types, and Tier-3 mixup cycles its weight over $\{0.1, 0.2, 0.35, 0.5, 0.65, 0.8, 0.9\}$ with random pairs (same-class pairs occur at chance rate). Masking uses five shapes (one block, random pixels, centre block, several blocks, border) at occluded fractions $\{0.3, 0.5, 0.7, 0.8\}$, filled with the dataset mean (on MNIST with 0, the background value). MNIST mixup is computed in pixel space and clipped to $[0, 1]$. *MNIST* (60k / 30k / 30k): Gaussian noise (σ=0.3), blur (5×5, σ=2), contrast 0.3, inversion, coloured background, coloured digits, salt-and-pepper (3%), pixelation; mixup and masking. *CIFAR-10* (50k / 25k / 25k): Gaussian noise (σ=0.15 in normalised space), blur (5×5, σ=2), contrast 0.3; mixup (clipped to $[0, 1]$ in normalised space) and masking. *Tiny ImageNet* (30k / 15k / 15k): Gaussian noise, blur ($\sigma \sim U(1, 2)$), contrast 0.5, JPEG (q=10), brightness $\times1.6$, shot noise; Tier 3 has seven equal parts: mixup, patch shuffling, cut-and-paste, masking, heavy noise, combined corruptions, and pixel shuffling within ViT patches. Tier-2 inputs are corrupted copies of validation images (App. B.8), and the Gaussian noise and blur are drawn again at training time. *VOC* (1,464 / 366 / 366 images): noise (re-drawn at every epoch), blur, contrast on the first 122 training images each; mixup with weights $\{0.2, 0.5, 0.8\}$ and single-block masking. The Tier-3 targets are not the teacher's MI: the pixel-wise maximum of the two source maps for mixup, and the unmasked image's map for masking. *SST-2* (67k / 17k / 17k): character-level typos (swaps, insertions, keyboard substitutions) and WordNet synonym replacement; Tier 3 is token masking at 30, 50 or 70%. The token masks used in training are drawn with a different seed from those on which the cached targets were computed. Retraining Stage 2 (held-out protocol) with the target masks changes little: dev variance ratio 7.58 instead of 7.60, and OOD AUROCs within 0.03. Apart from the re-drawn corruptions and masks above, the VOC Tier-3 targets, and the VOC Stage-1 member, the teacher is run once on every stored input and the targets are cached, and it is not queried during student training.

C.2 Distance-Aware Baselines: Mechanism Summary

Table 22 summarises the mechanism and the architectural constraint of the four distance-aware methods mentioned in §2 (DUQ, SNGP, DDU, DUE); we run SNGP, DDU and DUE as baselines.

Table 22: Distance-aware single-pass UQ methods: mechanism and architectural constraint. None of the four is trained on posterior samples; each score is a feature-space proxy.

Method	Mechanism	Architectural constraint
DUQ (van Amersfoort et al., 2020)	Per-class RBF centroids; distance to nearest centroid is the uncertainty score.	Two-sided gradient penalty during training; expensive at high resolution.
SNGP (Liu et al., 2020a; 2023)	Random-feature approximate GP head; posterior variance is the uncertainty score.	`SpectralBatchNorm2d` on every hidden layer (per-layer Lipschitz bound).
DDU (Mukhoti et al., 2023)	Post-hoc per-class GDA on penultimate features; Mahalanobis distance is the EU score.	Spectral normalisation on backbone; single Gaussian per class.
DUE (van Amersfoort et al., 2022)	Inducing-point approximate GP head on top of a bi-Lipschitz feature extractor.	Bi-Lipschitz regulariser; inducing-point bottleneck at large class counts.

C.3 Win/Tie/Loss Summary Against Each Baseline Family

For every baseline and benchmark we compare the student's score with the baseline's on each OOD set (AUROC), on AURC and on misclassification AUROC. A comparison is a win or a loss when the 95% paired-bootstrap interval of the difference ($1{,}000$ resamples) excludes zero, and a tie otherwise.

Table 23: Student vs. each baseline on MNIST, CIFAR-10 and Tiny ImageNet: *wins / ties / losses* decided by 95% paired-bootstrap intervals. OOD AUROC is counted per OOD set ($5 + 6 + 7$ sets); AURC and misclassification AUROC once per benchmark. Benchmarks on which a baseline is not run are excluded (App. D.8).

Baseline	OOD AUROC	AURC	Miscls. AUROC	Benchmarks
Multi-pass				
Teacher	3 / 5 / 10	0 / 0 / 3	0 / 3 / 0	3
MC Dropout	15 / 2 / 1	3 / 0 / 0	1 / 2 / 0	3
Laplace (LLLA)	11 / 0 / 7	3 / 0 / 0	1 / 2 / 0	3
SGLD	18 / 0 / 0	3 / 0 / 0	2 / 1 / 0	3
SWAG	12 / 0 / 6	1 / 1 / 1	1 / 0 / 2	3
Single-pass				
EDL	14 / 2 / 2	3 / 0 / 0	0 / 3 / 0	3
SNGP	7 / 0 / 4	1 / 1 / 0	0 / 2 / 0	2
DDU	10 / 2 / 6	3 / 0 / 0	3 / 0 / 0	3
DUE	6 / 0 / 0	1 / 0 / 0	1 / 0 / 0	1
DEUP	18 / 0 / 0	3 / 0 / 0	2 / 1 / 0	3
Energy (member)	11 / 0 / 7	3 / 0 / 0	3 / 0 / 0	3
ViM (member)	9 / 2 / 7	3 / 0 / 0	3 / 0 / 0	3
ViM (student backbone)	5 / 2 / 11	3 / 0 / 0	3 / 0 / 0	3

SNGP: MNIST and CIFAR-10; DUE: CIFAR-10.

D Full Vision Results

This section gives the per-baseline results behind the main-text tables, by evaluation axis: fidelity, OOD detection, selective prediction, calibration and throughput. The baseline settings are listed in §D.8.

D.1 Distillation Fidelity: Correlation with Teacher MI

Table 24 reports the Pearson correlation between each method's uncertainty score and the teacher's MI on clean test data. This measures agreement with the teacher, not uncertainty quality; only the student is trained on this target.

Table 24: Fidelity: Pearson correlation ↑ with the teacher's MI, per baseline, on the three vision benchmarks. Bold marks the best non-teacher cell per benchmark, underline the second-best. Bracketed values mark ours. "—": not run (App. D.8).

Method	MNIST ↑	CIFAR-10 ↑	Tiny ImageNet ↑
Teacher	1.000	1.000	1.000
Ours	[**0.818**]	[**0.815**]	[**0.842**]
MC Dropout	0.698	<u>0.633</u>	0.786
EDL	0.747	0.597	0.695
Laplace (LLLA)	<u>0.789</u>	0.600	<u>0.818</u>
SGLD	0.766	0.482	0.439
SWAG	0.747	0.527	0.772
DDU	0.414	0.514	0.706
SNGP	0.755	0.562	—
DUE	—	0.501	—
DEUP (Lahlou et al., 2023)	0.610	0.603	0.467
Energy (member)	0.534	0.609	0.663
Δ Student vs. best non-teacher	+0.029	+0.182	+0.024

D.2 Fidelity Under Distribution Shift

Table 25 reports the Pearson and Spearman correlation between the student's score and the teacher's MI on Tiny ImageNet, per corruption type.

Table 25: Fidelity under distribution shift (Tiny ImageNet, $K=16$): Pearson and Spearman correlation between the student's score and the teacher's MI, per corruption type.

Corruption	Pearson r ↑	Spearman ρ ↑
Clean	0.842	0.885
Gaussian noise	0.872	0.899
Gaussian blur	0.880	0.904
Low contrast	0.852	0.862
JPEG compression	0.889	0.918
Brightness	0.884	0.907
Shot noise	0.830	0.864

D.3 Full OOD Detection Results

Tables 26 and 28 report OOD-detection AUROC for every baseline. "Laplace (LLLA)" is KFAC last-layer Laplace.

AUPR and FPR@95. Table 27 reports AUPR-In, AUPR-Out and FPR@95 for the teacher and the student.

Post-hoc OOD scoring baselines. Table 28 reports two post-hoc detectors: Energy (Liu et al., 2020b) (negative free-energy $-T \log \sum_c e^{z_c/T}$) and ViM (Wang et al., 2022) (residual norm in the principal-subspace null space). Both are designed as OOD scores; Tables 2 and 7 also score them on selective prediction.

Table 26: OOD detection AUROC ↑ of the multi-pass Bayesian and single-pass deterministic baselines. **Bold** marks the best non-teacher cell and underline the second-best within each row.

(a) Tiny ImageNet (ID = Tiny ImageNet val, DeiT-Small + LoRA, $K=16$)

OOD Dataset ↑	Teacher	Ours	MC Dropout	Laplace (LLLA)	SGLD	SWAG	EDL	DDU
SVHN	0.980	0.898	0.761	0.913	0.641	**0.946**	0.774	0.839
CIFAR-10	0.872	0.831	0.812	0.864	0.635	0.838	0.828	**0.880**
CIFAR-100	0.867	0.825	0.784	**0.857**	0.642	0.837	0.802	0.856
STL10	0.812	0.789	0.691	**0.824**	0.556	0.746	0.726	0.796
DTD	0.909	0.899	0.743	**0.922**	0.666	0.873	0.791	0.914
FashionMNIST	0.945	0.895	0.799	**0.937**	0.784	0.911	0.825	0.932
MNIST	0.895	0.887	0.713	0.917	0.685	0.916	0.793	**0.930**

(b) CIFAR-10 (ID = CIFAR-10 test, ResNet-18, $K=16$)

OOD Dataset ↑	Teacher	Ours	MC Dropout	Laplace (LLLA)	SGLD	SWAG	EDL	SNGP	DDU	DUE
SVHN	0.981	**0.984**	0.956	0.942	0.878	0.924	0.927	0.923	0.964	0.943
CIFAR-100	0.925	**0.920**	0.893	0.853	0.797	0.878	0.900	0.890	0.844	0.862
MNIST	0.935	0.943	0.942	0.883	0.903	0.891	0.942	**0.960**	0.887	0.898
FashionMNIST	0.945	0.942	0.910	0.938	0.914	0.895	0.957	**0.963**	0.924	0.920
STL10	0.672	0.668	0.668	0.637	0.620	**0.686**	0.677	0.681	0.629	0.653
DTD	0.958	**0.948**	0.940	0.921	0.865	0.907	0.929	0.919	0.947	0.920

SWAG recomputes the BatchNorm statistics for each weight sample (CIFAR-10, MNIST). DUE is trained for 200 epochs (95.3% accuracy).

(c) MNIST (ID = MNIST test, 4-layer CNN, $K=16$)

OOD Dataset ↑	Teacher	Ours	MC Dropout	Laplace (LLLA)	SGLD	SWAG	EDL	SNGP	DDU
FashionMNIST	0.999	**0.999**	0.998	0.933	0.917	0.886	0.910	0.970	0.991
Omniglot	1.000	**1.000**	**1.000**	0.799	0.720	0.923	0.699	0.980	**1.000**
EMNIST-Letters	0.972	0.958	**0.989**	0.942	0.920	0.942	0.934	0.961	0.914
CIFAR-10	1.000	**1.000**	0.998	0.853	0.754	0.235	0.600	0.976	**1.000**
SVHN	1.000	**1.000**	0.998	0.989	0.937	0.296	0.883	0.979	0.999

Table 27: AUPR-In ↑, AUPR-Out ↑ and FPR@95 ↓ of the teacher and the student. On MNIST most cells are saturated for both models; on Tiny ImageNet the student trails the teacher on far-OOD sets, as in Table 26.

		Teacher			Ours		
Benchmark	OOD set	AUPR-In ↑	AUPR-Out ↑	FPR@95 ↓	AUPR-In ↑	AUPR-Out ↑	FPR@95 ↓
MNIST	FashionMNIST	0.998	0.999	0.003	0.999	0.999	0.007
	Omniglot	1.000	1.000	0.000	1.000	1.000	0.000
	SVHN	1.000	1.000	0.000	1.000	1.000	0.000
	EMNIST-Letters	0.945	0.985	0.135	0.915	0.978	0.190
CIFAR-10	SVHN	0.967	0.990	0.081	0.969	0.992	0.081
	CIFAR-100	0.929	0.908	0.248	0.926	0.907	0.282
	MNIST	0.951	0.893	0.166	0.955	0.917	0.196
	FashionMNIST	0.958	0.910	0.148	0.948	0.932	0.219
	STL10	0.673	0.638	0.905	0.671	0.635	0.910
	DTD	0.992	0.796	0.151	0.990	0.725	0.169
Tiny ImageNet	SVHN	0.991	0.955	0.093	0.957	0.703	0.241
	CIFAR-100	0.924	0.750	0.482	0.911	0.619	0.500
	DTD	0.982	0.617	0.279	0.980	0.535	0.299
	MNIST	0.957	0.690	0.232	0.952	0.668	0.261

D.4 Full Selective Prediction Results

Table 29 reports per-baseline selective-prediction metrics: AURC (area under the risk-coverage curve), E-AURC (excess AURC over the oracle ranking), and accuracy at the 90%/80% coverage operating points.

Table 28: Post-hoc OOD scores (AUROC ↑): Energy and ViM on the backbone of one teacher member (member 0), and ViM on the student's backbone. ViM uses a principal subspace of half the feature dimension.

Benchmark	OOD Dataset	Energy (member)	ViM (member)	ViM (student)
MNIST	FashionMNIST	0.748	0.984	1.000
	Omniglot	0.253	1.000	1.000
	EMNIST-Letters	0.891	0.894	0.975
	CIFAR-10	0.188	0.998	1.000
	SVHN	0.539	0.996	1.000
CIFAR-10	SVHN	0.924	0.973	0.973
	CIFAR-100	0.823	0.855	0.904
	MNIST	0.857	0.910	0.944
	FashionMNIST	0.926	0.939	0.955
	STL10	0.620	0.642	0.696
	DTD	0.856	0.953	0.958
Tiny ImageNet	SVHN	0.965	0.916	0.853
	CIFAR-10	0.842	0.893	0.880
	CIFAR-100	0.847	0.869	0.859
	STL10	0.862	0.864	0.867
	DTD	0.942	0.949	0.959
	FashionMNIST	0.939	0.946	0.941
	MNIST	0.919	0.956	0.900

D.5 Calibration

Table 30 summarises student-vs-teacher calibration on clean test data (ECE-15, NLL, Brier); Table 31 reports the per-baseline numbers.

D.6 Calibration under CIFAR-10-C corruption (full 15×5)

Table 32 reports per-severity mean ECE-15 ($\times 100$) averaged over the 15 standard CIFAR-10-C corruption types.

D.7 Inference Throughput

Table 33 reports throughput for all six benchmarks, measured with one protocol on one GPU after warm-up (SST-2 with batch size 32 and sequence length 128); LoRA adapters are merged into the backbone weights before timing. The student runs at the speed of one member, so the speed-up is close to K; on SST-2 the DistilBERT student is about twice as fast as a BERT-base member. Table 35 compares CIFAR-10 single-pass baselines.

D.8 Baseline Settings

Table 34 lists the benchmarks on which each baseline is run and its settings.

Single-pass baseline throughput (CIFAR-10). Table 35 reports wall-clock throughput for each single-pass deterministic baseline and the multi-pass references on CIFAR-10 (ResNet-18 unless noted, batch 256, FP32, RTX 4090; `torch.compile(fullgraph=True)`, 100 warm-up iterations).

E Ablations and Diagnostics

This section gives the full ablation tables (§E.1–§E.2) behind Table 5 and the diagnostic figures (§E.3).

Table 29: Selective prediction across baselines. AURC ↓, E-AURC ↓ (excess AURC over the oracle ranking), Acc@90% ↑, Acc@80% ↑. **Bold** / underline mark the best / second-best non-teacher cell per column.

(a) MNIST (K=16)

Method	AURC ↓	E-AURC ↓	Acc@90% ↑	Acc@80% ↑
Teacher	0.000111	0.000100	99.99%	100.00%
Ours	**0.000147**	0.000135	99.99%	**100.00%**
MC Dropout	0.000333	0.000299	99.93%	99.99%
EDL	0.000216	0.000179	99.97%	**100.00%**
SGLD	0.000284	0.000245	99.96%	**100.00%**
Laplace (LLLA)	0.000252	0.000220	99.97%	**100.00%**
DDU	0.001362	0.001330	99.62%	99.75%
SWAG	0.000152	**0.000128**	**100.00%**	**100.00%**
SNGP	0.000168	0.000152	99.98%	**100.00%**

(b) CIFAR-10 (K=16)

Method	AURC ↓	E-AURC ↓	Acc@90% ↑	Acc@80% ↑
Teacher	0.002346	0.001879	99.43%	99.80%
Ours	**0.003427**	**0.002612**	**99.04%**	**99.72%**
MC Dropout	0.007044	0.004971	97.27%	98.98%
EDL	0.006403	0.004250	97.98%	99.28%
Laplace (LLLA)	0.009445	0.007940	98.04%	98.88%
SGLD	0.017505	0.016053	97.69%	98.31%
DDU	0.009850	0.008399	97.07%	98.34%
SWAG	0.006495	0.004600	97.70%	99.33%
SNGP	0.006627	0.004639	97.92%	99.12%

(c) Tiny ImageNet (K=16)

Method	AURC ↓	E-AURC ↓	Acc@90% ↑	Acc@80% ↑
Teacher	0.0224	0.0151	92.27%	95.66%
Ours	0.0243	0.0161	91.70%	95.15%
MC Dropout	0.0297	0.0188	90.00%	93.47%
Laplace (LLLA)	0.0282	0.0187	91.46%	94.46%
EDL	0.0315	0.0191	89.76%	93.83%
SGLD	0.0886	0.0499	76.84%	81.50%
DDU	0.0385	0.0297	90.89%	93.51%
SWAG	**0.0204**	**0.0120**	**92.37%**	**96.03%**

Table 30: Calibration summary on clean test data: teacher vs. student. **Bold** marks the lower (better) value in each metric pair, per benchmark. Per-baseline numbers in Table 31.

Benchmark	Teacher ECE ↓	Student ECE ↓	Teacher NLL ↓	Student NLL ↓	Teacher Brier ↓	Student Brier ↓
MNIST	0.0377	**0.0372**	**0.0510**	0.0519	**0.0122**	0.0130
CIFAR-10	0.0294	**0.0195**	**0.1153**	0.1425	**0.0487**	0.0639
Tiny ImageNet	0.0763	**0.0484**	**0.4862**	0.5139	**0.1778**	0.1832

E.1 Full Ablation Tables

Tables 36–38 cover three design choices on CIFAR-10 (K=16): two-stage vs. joint training, the Stage-2 loss, and the Stage-2 auxiliary set.

Table 31: Per-baseline calibration on the clean test set: accuracy, ECE-15 ↓ and NLL ↓. ‡: LLLA with a temperature fitted on a random 20% of the test set; its ECE and NLL are computed on the remaining 80%. †: DUE with its official recipe. "—": not run. Baseline settings are in App. D.8.

Benchmark	Metric	Teacher	Ours	MC Dropout	LLLA‡	SGLD	SWAG	SNGP	DDU	EDL	DUE
MNIST	Acc%	99.54	99.51	99.17	99.20	99.12	99.31	99.44	99.20	99.14	—
	ECE-15	0.0377	0.0372	0.0221	0.0014‡	0.0202	0.0121	0.0419	0.0386	0.0913	—
	NLL	0.0510	0.0519	0.0434	0.0251‡	0.0426	0.0297	0.0579	0.0612	0.1247	—
CIFAR-10	Acc%	96.96	95.99	93.63	94.56	94.66	93.91	93.76	94.66	93.51	95.32†
	ECE-15	0.0294	0.0195	0.0085	0.0195‡	0.0216	0.0251	0.0208	0.0259	0.0471	0.0095†
	NLL	0.1153	0.1425	0.2006	0.2191‡	0.2131	0.2176	0.2607	0.2172	0.2994	0.7313†
Tiny ImageNet	Acc%	88.13	87.50	85.60	86.51	73.50	87.33	—	87.01	84.70	—
	ECE-15	0.0763	0.0484	0.0359	0.0103‡	0.0894	0.0240	—	0.0321	0.1955	—
	NLL	0.4862	0.5139	0.5229	0.4786‡	1.1355	0.4688	—	0.4956	1.4409	—

Table 32: Mean ECE-15 ($\times 100$) ↓ averaged across the 15 standard CIFAR-10-C corruption types, by severity. **Bold** marks the best score per column. The student has the lowest ECE on clean test data; the teacher's advantage grows with severity.

Method	sev=0 (clean) ↓	sev=1 ↓	sev=2 ↓	sev=3 ↓	sev=4 ↓	sev=5 ↓
Ours	**1.95**	4.16	7.01	10.70	15.10	22.29
MSP (single member)	2.59	**3.21**	4.97	7.87	12.39	20.27
Teacher (K=16)	2.94	3.76	**4.10**	**5.22**	**6.28**	**8.56**

E.2 Sensitivity Sweeps (β and K)

Table 41 sweeps the Stage-2 ranking weight β on a fixed Stage-1 backbone: Pearson peaks near β=0.25 and Spearman rises monotonically with β. Table 42 sweeps the teacher size $K \in \{3, 5, 10, 16\}$, retraining Stage 1 for each K. Fidelity rises monotonically with K (Pearson $0.63 \rightarrow 0.81$), with most of the gain by K=10; mean OOD AUROC and student calibration change little beyond K=5.

E.3 Diagnostic Figures

This subsection collects the diagnostic figures: the per-input scatter (Figure 2) and density (Figure 3) of the student's score against the teacher's MI; reliability diagrams (Figure 4); and a PCA-basis diagnostic (Figure 5).

E.3.1 Per-input scatter and density

Figure 2 shows the per-input scatter of the student's score against the teacher's MI; Figure 3 shows the density of the student's score by ID/OOD source.

E.3.2 Reliability diagrams

Figure 4 is the diagram view of the ECE numbers in Table 30.

E.3.3 PCA-basis diagnostic (heuristic)

Figure 5 shows, for each principal component of the head's input features, the ratio of the student's to the teacher's variance along that component. It is not a test of Theorem 1: coordinate-wise shrinkage holds only in the kernel's eigenbasis, and in another basis a contracting filter can enlarge individual components. The test in the head's empirical NTK basis is in App. B.4.

Table 33: Inference throughput (samples/s) of the teacher and the student, one protocol for all benchmarks.

Benchmark	Teacher	Ours	Speed-up
MNIST (K=16)	40,447	627,661	15.5×
CIFAR-10 (K=16)	2,073	33,088	16.0×
CIFAR-100 (K=5)	6,634	33,095	5.0×
Tiny ImageNet (K=16)	430	6,916	16.1×
VOC (K=5, images)	35	171	4.9×
SST-2 (K=5)	280	2,785	9.9×

Table 34: Baseline settings on the three vision benchmarks.

Method	Benchmarks	Settings
MC Dropout	MNIST, CIFAR-10, Tiny ImageNet	16 stochastic passes; BatchNorm uses test-batch statistics. CIFAR-10: Dropout2d (p=0.1) after residual stages 1–3, fine-tuned for 20 epochs from one teacher member.
SNGP	MNIST, CIFAR-10	`SpectralBatchNorm2d` on every hidden layer.
DUE	CIFAR-10	Official recipe: WideResNet-28-10 with `SpectralBatchNorm2d`, long schedule; 95.32% accuracy, fidelity (Pearson) 0.501. Table 35 uses this backbone for DUE.
EDL	MNIST, CIFAR-10, Tiny ImageNet	Evidential loss of Sensoy et al. (2018) with KL weight annealed from 0 to 1 over the first 10 epochs. Tiny ImageNet: linear evidential head on one teacher member's frozen features, exponential evidence, log loss, 50 epochs, final epoch, three seeds.
DDU	MNIST, CIFAR-10, Tiny ImageNet	Per-class GDA on penultimate features. On backbones without spectral normalisation this is identical to Mahalanobis (Lee et al., 2018); reported as a single row.
SWAG	MNIST, CIFAR-10, Tiny ImageNet	Variance scale 0.5 (Maddox et al., 2019), 20 weight samples, BatchNorm statistics recomputed for each weight sample.
DEUP	MNIST, CIFAR-10, Tiny ImageNet	MLP error predictor on one teacher member's penultimate features, trained in log space on the out-of-fold negative log-likelihood of the true class (two-fold split of the training set); its output is the score.
LLLA	MNIST, CIFAR-10, Tiny ImageNet	KFAC last-layer Laplace on one teacher member, 16 posterior samples of the predictive distribution. Post-hoc temperature fitted on a random 20% of the test set (0.47 / 0.86 / 0.74 on MNIST / CIFAR-10 / Tiny ImageNet).
Energy / ViM	MNIST, CIFAR-10, Tiny ImageNet	Post-hoc OOD scores, also scored on selective prediction. Computed on one teacher member's backbone in the per-set tables; §5.3 and App. B.12 also report them on the student's backbone. ViM uses a principal subspace of half the feature dimension, as in the official code.

Table 35: Inference throughput on CIFAR-10 (ResNet-18 unless noted, batch 256, FP32), measured in one run with `torch.compile` so that all methods share the same setting; absolute numbers therefore differ slightly from Table 33. **Bold** / underline mark the best / second-best within the single-pass block.

Method	samples/s ↑	relative to student ↑
Multi-pass		
Teacher (K=16)	2,082	0.063×
MC Dropout (T=16)	2,009	0.061×
SGLD (last-layer, T=16)	32,511	0.980×
Single forward pass		
Single ResNet-18	32,845	0.990×
Ours	**33,163**	**1.000×**
EDL	33,094	0.998×
SNGP	20,921	0.631×
DDU ≡ Mahalanobis	25,537	0.770×
DUE (WideResNet-28-10, official)	4,160	0.125×
Energy (post-hoc)	33,132	0.999×
ViM (post-hoc)	33,031	0.996×

Table 36: Two-stage vs. joint training (CIFAR-10, K=16). Joint training is lower on all four metrics. "Mean OOD AUROC" is the average over six OOD sets (CIFAR-100, DTD, FashionMNIST, MNIST, STL10, SVHN); per-set numbers in Table 39. **Bold** marks the better of the two strategies per column.

Strategy	Acc ↑	Pearson r ↑	Spearman ρ ↑	Mean OOD AUROC ↑
Two-stage (ours)	**95.99%**	**0.815**	**0.799**	**0.901**
Joint (γ=1.0)	95.07%	0.646	0.627	0.853

Table 37: Stage-2 loss ablation (CIFAR-10, K=16). **Bold** marks the best score per column; underline the second-best. All four variants lie within 0.01 on mean OOD AUROC. Per-set OOD numbers in Table 39.

Loss	Pearson r ↑	Spearman ρ ↑	Mean OOD AUROC ↑
Plain MSE	0.814	0.644	**0.909**
Log-MSE	**0.824**	0.650	0.907
Ranking only	0.704	**0.818**	0.901
Log-MSE + Ranking (ours)	0.815	0.799	0.901

Table 38: Stage-2 auxiliary-set ablation (CIFAR-10, K=16). **Bold** marks the best score per column; underline the second-best. Adding Tier 2 raises mean OOD AUROC by +0.021; adding Tier 3 changes it by −0.004. Per-set numbers in Table 39.

Auxiliary set	Tier mix	Pearson r ↑	Spearman ρ ↑	Mean OOD AUROC ↑
Tier 1 only	100% clean	0.800	**0.807**	0.884
No Tier 3	67% / 33%	0.814	0.807	**0.905**
Tiers 1–3 (ours)	50% / 25% / 25%	**0.815**	0.799	0.901

Table 39: Per-OOD-set breakdown of the mean OOD AUROC in Tables 36–38 and Table 5 (CIFAR-10, $K=16$). The last column is the mean. **Bold** marks the best cell per OOD set across the six variants (ranked by raw values), underline the second-best.

Variant	C-100 ↑	DTD ↑	F-MNIST ↑	MNIST ↑	STL10 ↑	SVHN ↑	Mean ↑
Ours	0.920	0.948	0.942	0.943	0.668	0.984	0.901
Joint training	0.871	0.860	0.878	0.914	0.665	0.929	0.853
Tier 1 only	0.904	0.920	0.928	0.943	0.665	0.945	0.884
Ranking only	**0.921**	0.949	0.935	0.945	0.668	0.986	0.901
Plain MSE	0.920	0.945	**0.954**	0.952	**0.695**	0.985	**0.909**
No Tier 3	0.916	**0.951**	0.944	**0.966**	0.667	**0.989**	0.905

Table 40: Per-severity breakdown of the CIFAR-10-C AUROC in Table 5: AUROC of clean CIFAR-10 test (10 000 images, negatives) vs. CIFAR-10-C (positives), per severity, averaged over the 19 corruption types. The last column is the mean over the 5 severities (the number in Table 5). **Bold** marks the best cell per severity across the six variants (ranked by raw values), underline the second-best.

Variant	$s=1$ ↑	$s=2$ ↑	$s=3$ ↑	$s=4$ ↑	$s=5$ ↑	Mean ↑
Ours	**0.606**	**0.671**	**0.717**	**0.766**	**0.833**	**0.719**
Joint training	0.592	0.644	0.685	0.727	0.784	0.686
Tier 1 only	0.598	0.657	0.697	0.739	0.801	0.698
Ranking only	0.605	0.669	0.714	0.762	0.830	0.716
Plain MSE	0.601	0.670	0.717	0.759	0.831	0.716
No Tier 3	0.606	0.669	0.713	0.760	0.828	0.715

Table 41: Stage-2 ranking weight β sweep on CIFAR-10 ($K=16$). Backbone fixed; accuracy and ECE-15 are constant. The sweep uses its own Stage-1 backbone (95.98% accuracy) and Stage-2 schedule (50 epochs, lr 5×10^{-3}, last epoch kept), so its $\beta=1$ row differs slightly from the main model (0.815 / 0.799). **Bold** marks the best per column, underline the second-best.

β	Acc% ↑	Pearson r ↑	Spearman ρ ↑	SVHN AUROC ↑
0.0	95.98	0.828	0.659	**0.984**
0.25	95.98	**0.830**	0.782	0.983
0.5	95.98	0.828	0.790	0.983
1.0	95.98	0.821	0.799	0.982
2.0	95.98	0.812	0.802	0.983
5.0	95.98	0.778	**0.807**	0.983

Table 42: Teacher size K sweep on CIFAR-10. Each row trains Stage 1 from scratch with a K-member teacher and a new Stage-2 head on the resulting backbone (the $K=16$ row is the main model). "Mean OOD AUROC" averages over six OOD sets (CIFAR-100, DTD, FashionMNIST, MNIST, STL10, SVHN). **Bold** marks the best per column, underline the second-best. Pearson rises by $+0.009$ from $K=10$ to $K=16$; from $K=5$ on, mean OOD AUROC and student ECE lie in 0.901–0.907 and 0.0180–0.0209. Teacher ECE-15 rises mildly with K.

K	Acc% ↑	Student ECE-15 ↓	Teacher ECE-15 ↓	Pearson r ↑	Spearman ρ ↑	Mean OOD AUROC ↑
3	95.78	**0.0158**	**0.0230**	0.631	0.645	0.897
5	95.60	0.0209	0.0258	0.717	0.717	**0.907**
10	95.84	0.0180	0.0281	0.806	0.790	0.901
16 (main)	**95.99**	0.0195	0.0294	**0.815**	**0.799**	0.901

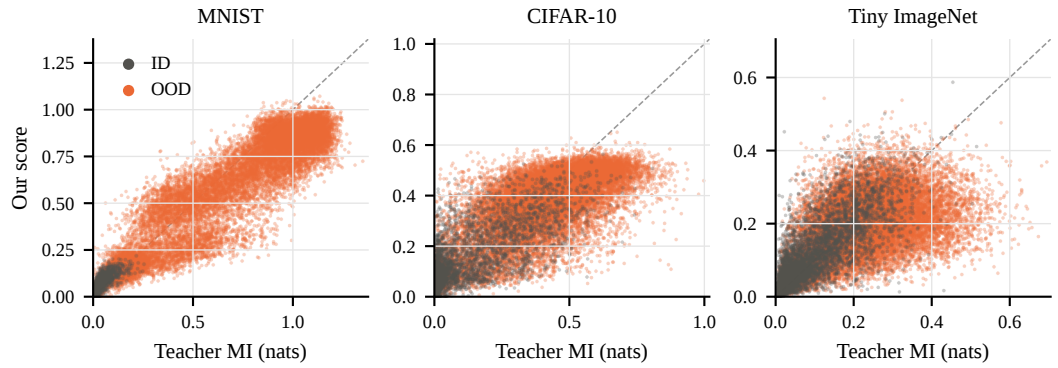

Figure 2: **Per-input scatter of our student's score against the teacher's MI** on MNIST, CIFAR-10 and Tiny ImageNet. Grey: all 10,000 ID test inputs; orange: a random subsample of 20,000 inputs from the pooled OOD sets. Dashed line: $y = x$. ID points concentrate near the origin.

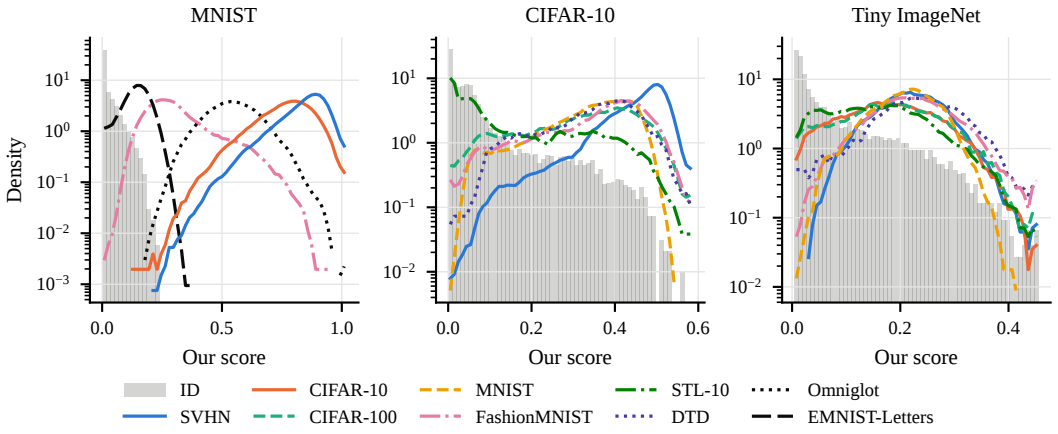

Figure 3: **Density of our student's score on ID and on each OOD set** (log scale; 60 bins over the 0.1–99.9% quantile range of each panel, with values outside it placed in the edge bins). ID is shown as a histogram; each OOD curve is a 3-point moving average of its 60-bin histogram. The overlap between the ID distribution and each OOD distribution determines the OOD AUROC in Table 26.

Algorithm 1 Offline training of our student. Benchmark-specific differences are listed in §3.4 and App. C.1.

Require: training set $\mathcal{D}$; teacher members $\{f_k\}_{k=1}^{K}$; student backbone and task head; uncertainty head g_ϕ

1: **Auxiliary set.** Build Tier 2 by corrupting inputs of $\mathcal{D}$ (e.g. Gaussian noise, blur, contrast; one fixed severity each) and Tier 3 by mixup of randomly paired inputs (mixing weight cycled over a fixed grid) and by large occlusions; Tier 1 is $\mathcal{D}$ itself. Store the generated inputs.
2: **Teacher pass (once).** For every input x in the three tiers, cache the averaged predictive distribution $\bar{p}(x)$ and $u_T(x)$ of Eq. (3).
3: **Stage 1.** Train backbone and task head on $\mathcal{D}$ with $\mathcal{L}_{\text{pred}}$ (Eq. 4).
4: **Stage 2.** Freeze backbone and task head. Train g_ϕ on $[\mathbf{h}(x) \,\|\, \sigma(z_S)]$ with $\mathcal{L}_{\text{unc}}$ (Eq. 5), on the shuffled union of the three tiers (50/25/25% of the set by default).
5: **Deploy** the student alone: one forward pass returns $\hat{y}$ and $\hat{u}(x) = g_\phi(\cdot)$.

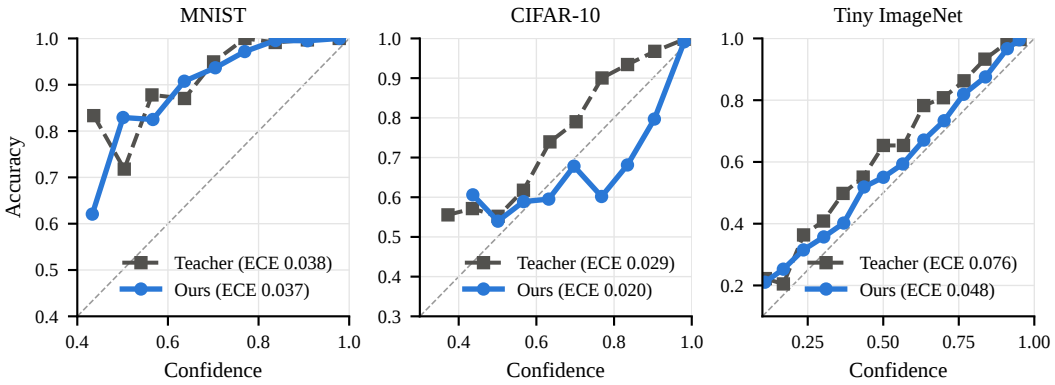

Figure 4: **Reliability diagrams** of the teacher and our student on each benchmark (15 equal-width confidence bins; bins with fewer than 20 inputs are omitted; axis ranges differ per panel). On CIFAR-10 and Tiny ImageNet the student's curve is closer to the diagonal than the teacher's (lower ECE in Table 30).

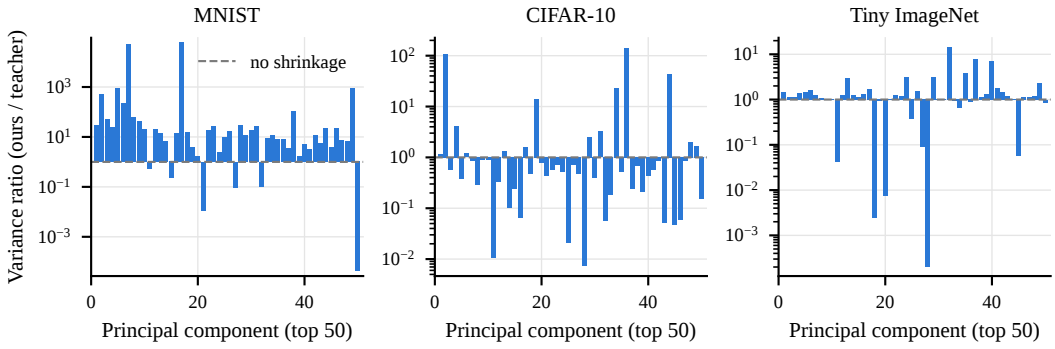

Figure 5: **Per-PCA-component variance ratio** (our student / teacher): for each of the top 50 principal directions of the head's input features, the squared projection of our student's score divided by that of the teacher's MI (log scale; bars start at ratio 1, marked by the dashed line). Heuristic diagnostic; the test of the analysis is in App. B.4.

F Cross-Domain and Deployment Results

This section reports the cross-domain stress tests (segmentation on PASCAL VOC 2012 §F.1, sentiment on SST-2 §F.2), the deployment-time experiments on CIFAR-10-C (suitability monitoring and cascade routing, §F.3), and repeated self-distillation (§F.4). All experiments use the same two-stage pipeline as the core vision benchmarks.

F.1 Semantic Segmentation on PASCAL VOC 2012

Setup

Dataset. PASCAL VOC 2012 augmented with SBD (Hariharan et al., 2011) gives $10{,}582$ training images and $1{,}449$ validation images across 21 semantic categories. We crop and resize all images to 512×512.

Teacher ensemble. We train $K=5$ SegFormer-B2 members (Xie et al., 2021), each fine-tuned from the ADE20K-pretrained checkpoint via LoRA (rank 16, $\alpha=32$, target: query and value projections in all attention layers). Member diversity is induced through augmentation strategy ($\in$ {scale, colour jitter, rotation}), label smoothing ($\in [0.0013, 0.0348]$), weight decay ($\in \{10^{-4}, 5\times10^{-4}, 10^{-3}\}$), and random seed. Only the LoRA and decode-head parameters of each member are stored; the decode head's BatchNorm running statistics are those of the pretrained checkpoint. Table 43 lists per-member configurations and validation mIoU. The members, the Stage-1 checkpoint and the Stage-2 checkpoint are all selected on the validation set, which is also the evaluation set.

Table 43: VOC: teacher member configurations, selected epoch (best validation mIoU), and mIoU of each member as used in the teacher.

Member	Seed	Augmentation	Label Smooth	Weight Decay	mIoU ↑	Selected epoch
0	42	scale	0.0013	10^{-4}	0.7849	105
1	43	colour jitter	0.0348	5×10^{-4}	0.7878	195
2	44	colour jitter	0.0271	10^{-4}	0.7848	145
3	45	colour jitter	0.0244	5×10^{-4}	0.7878	195
4	46	rotation	0.0200	10^{-3}	0.7963	190
Mean over members					0.7883	

Student model. The student uses the same SegFormer-B2 backbone as its teachers. The uncertainty head g_ϕ operates on the 512-channel feature maps output by the backbone's hierarchical encoder (before the MLP decoder head), applying a 1×1 convolutional head:

$$g_\phi : \mathbb{R}^{H'\times W'\times 512} \xrightarrow{\text{Conv}(512\to128)\to\text{BN}\to\text{ReLU}} \mathbb{R}^{H'\times W'\times 128} \xrightarrow{\text{Conv}(128\to1)\to\text{Softplus}} \mathbb{R}^{H'\times W'},$$

producing a per-pixel epistemic uncertainty map that is bilinearly upsampled to the full 512×512 resolution. The head uses no global pooling, so it keeps the spatial resolution.

Stage 1 training. We distil the segmentation output of one teacher member (member 0, run online) with the pixel-wise KD objective, $\alpha=0.7$ and $\tau=4.0$. Optimiser: AdamW, lr$=6\times10^{-5}$, weight decay 0.01, polynomial decay ($p=0.9$) over a 40K-iteration schedule; training stopped at 22.9K iterations and we use the checkpoint with the best validation mIoU (step 20,130); batch size 8. The Stage-1 target is therefore a single member, not the ensemble.

Stage 2 training. The uncertainty head is trained for 73 epochs (20K iterations) on the frozen SegFormer-B2 + LoRA backbone. Adam (lr$=10^{-3}$, cosine schedule), batch size 8, rank weight $\beta=1.0$, ranking margin 0.05 on 512 random pixel pairs per batch. The auxiliary set has three tiers: clean VOC images (Tier 1), corrupted images (Tier 2: Gaussian noise, blur, contrast), and mixup and single-block masking (Tier 3), in proportions $67/17/17\%$. The Tier-3 targets are derived from the teacher's maps of the source images rather than computed by the teacher (App. C.1).

STAGE 1 RESULTS

Table 44: VOC Stage 1: segmentation accuracy (mIoU computed per image and averaged over the 1,449 validation images). Per-member mean: arithmetic mean of the K member mIoUs (Table 43). Per-pixel softmax-avg: the ensemble prediction obtained by averaging the members' per-pixel softmax outputs, i.e. the teacher.

Model	mIoU ↑
Teacher (per-member mean)	0.7883
Teacher (per-pixel softmax-avg)	0.8017
Ours	0.8056
Δ (vs. per-member mean)	+0.0173
Δ (vs. softmax-avg)	+0.0039

Our student's mIoU (0.806) is above the members' mean (0.788) and the ensemble's (0.802), at single-pass cost. With dataset-level mIoU (per-class IoU accumulated over all validation images) the ensemble is slightly ahead (0.798 vs. 0.793).

STAGE 2 RESULTS

Table 45: VOC Stage 2: per-pixel evaluation of the student on the 1,449-image VOC validation set.

Metric	Value
mIoU ↑	0.8056
Pixel accuracy ↑	95.02%
ECE-15 (per-pixel) ↓	0.0123
NLL ↓	0.1798
Per-pixel EU Pearson r (vs. teacher MI) ↑	0.201
Per-pixel EU Spearman ρ (vs. teacher MI) ↑	**0.559**
Boundary mean EU	0.0599
Interior mean EU	0.0585
Boundary/Interior ratio	1.02
AURC (selective segmentation) ↓	0.0437
Acc. @ 50% coverage ↑	95.65%
Acc. @ 90% coverage ↑	95.60%
Throughput (RTX 4090, batch 4) ↑	**171** img/s

At the pixel level, the Spearman correlation between the student's score and the teacher's MI is ρ=0.559 and the Pearson correlation is 0.201. The gap reflects a heavy-tailed pixel-level MI distribution, so Spearman is the more informative measure of ranking quality here. Per-pixel ECE is 0.012. Throughput is 4.9× that of the teacher at the same resolution (App. D.7). We did not evaluate pixel-level OOD detection (e.g. RoadAnomaly, FishyScapes).

F.2 SENTIMENT CLASSIFICATION ON SST-2

SETUP

Dataset. SST-2 (Socher et al., 2013) provides 67,349 training sentences and 872 development sentences for binary sentiment classification. For OOD evaluation we use five text datasets that differ from SST-2 in domain and style: IMDB (movie reviews), Yelp (business reviews), Amazon (product reviews), AG News (news topics), and 20 Newsgroups (20NG).

Teacher ensemble. We train K=5 BERT-base-uncased members via LoRA (rank 8, α=16, target modules: all attention projections q, k, v, o). Member diversity is induced through random seed, LoRA initialisation scale, label smoothing, attention dropout, and learning rate. Table 46 lists per-member configurations.

Table 46: SST-2: teacher member configurations.

Member	Seed	Init Scale	Label Smooth	Attn Drop	LR	WD
0	42	1.009	0.00	0.15	1×10^{-4}	0.01
1	43	0.522	0.05	0.10	2×10^{-4}	0.01
2	44	0.783	0.05	0.10	1×10^{-4}	0.01
3	45	0.674	0.02	0.15	1×10^{-4}	0.01
4	46	1.327	0.02	0.10	3×10^{-4}	0.01

Student model. The student uses DistilBERT-base-uncased as its backbone (6 layers, 768 hidden). The uncertainty head receives the concatenation of the `[CLS]` embedding and the detached softmax output:

$$g_\phi(x) = \text{Softplus}(W_2\,\text{ReLU}(W_1\,[\mathbf{h}_{\text{CLS}}(x) \,\|\, \sigma(z_S(x))] + b_1) + b_2)\,,$$

with $W_1 \in \mathbb{R}^{256\times770}$, $W_2 \in \mathbb{R}^{1\times256}$, and a dropout layer ($p$=0.2) before W_2.

Stage 1 training. KD objective with α=0.5, τ=2.0; AdamW, lr=2×10^{-5}, weight decay 0.05, cosine schedule with 10% warmup, 50 epochs with early stopping, batch size 32, gradient clipping at 1.0, layer-wise LR decay (LLRD) 0.85. Text augmentation: random word deletion (10%) and swap (5%).

Stage 2 training. Uncertainty head only; Adam, lr=2×10^{-4}, cosine schedule, up to 200 epochs, batch size 32, rank weight β=1.0, 256 rank pairs, rank margin 0.05. Auxiliary set: clean SST-2 train (Tier 1, 67%), character/word-perturbed SST-2 (Tier 2, 17%), and token-masked SST-2 at 30/50/70% masking (Tier 3, 17%).

RESULTS

Table 47: SST-2: accuracy and calibration. Acc ↑, ECE ↓, NLL ↓, Brier ↓.

Model	Acc ↑	ECE-15 ↓	NLL ↓	Brier ↓
Teacher	91.51%	**0.0179**	**0.2122**	**0.1224**
Ours	**92.09%**	0.0251	0.2164	0.1272

Predictive accuracy and calibration. At single-pass cost, the student matches the teacher's accuracy (92.09% vs. 91.51%), with NLL and Brier within 0.005 of the teacher and ECE within 0.007.

Table 48: SST-2: fidelity (correlation of the student's score with the teacher's MI) on clean and perturbed text. **Bold** marks the higher value per column.

Split	Pearson r ↑	Spearman ρ ↑
Clean SST-2 dev	**0.486**	**0.577**
Char-perturbed SST-2 dev	0.434	0.532

Fidelity. Fidelity is moderate on both splits.

SST-2 baselines. We evaluate five baselines: MC Dropout, SGLD, SWAG, EDL and last-layer Laplace (LLLA, diagonal). The baselines use the BERT-base + LoRA teacher backbone; the student uses DistilBERT. Results are in Table 49.

The student's accuracy (92.09%) is on par with the K=5 teacher's (91.51%). The sampling baselines (MC Dropout, SGLD, SWAG) and LLLA correlate more closely with the teacher's MI, at higher inference cost. LLLA also has lower AURC (0.0194 vs. 0.0208) and higher mean OOD AUROC (0.663 vs. 0.609) than our student. The student's score has higher OOD AUROC than the teacher's

Table 49: SST-2 baselines on the BERT-base + LoRA teacher backbone, and the DistilBERT student. Fidelity: Pearson correlation with the $K=5$ teacher's MI. **Bold** / underline mark the best / second-best non-teacher cell per column.

Method	Acc% ↑	ECE-15 ↓	Fidelity (Pearson) ↑
Teacher ($K=5$)	91.51	0.0179	1.000
Ours	**92.09**	0.0251	0.486
MC Dropout ($T=16$)	90.94	**0.0158**	0.528
EDL	91.74	0.0601	0.437
SGLD (last-layer)	91.51	0.0166	0.589
SWAG (scale 0.5)	90.94	0.0380	0.591
LLLA (diagonal)	91.97	0.0284	**0.648**

on all five OOD sets (mean $+0.059$; Table 50) and 33% lower AURC (Table 51). The student was selected on the SST-2 dev set, which is also the evaluation set, so these gains may be optimistic. The analysis of §4 does not explain them: the student's variance exceeds the teacher's on every split, also with a same-backbone student (App. B.4).

Table 50: SST-2: OOD detection AUROC ↑ (ID = SST-2 dev).

OOD Dataset	Teacher ↑	Ours ↑
IMDB	0.477	**0.522**
Yelp	0.532	**0.600**
Amazon	0.501	**0.579**
AG News	0.663	**0.743**
20NG	0.578	**0.600**

OOD detection. Both scores are much lower than on the vision benchmarks, in line with results for ensemble uncertainty on short-text binary classification (Malinin & Gales, 2021).

Table 51: SST-2: selective prediction. AURC ↓, Acc@90% ↑, Acc@80% ↑.

Method	AURC ↓	Acc@90% ↑	Acc@80% ↑
Teacher	0.0309	94.64%	95.84%
Ours	**0.0208**	94.13%	96.41%
Oracle	0.0032	100.0%	100.0%

Selective prediction. The student's score lowers AURC by 33% relative to the teacher's (0.0208 vs. 0.0309).

Inference throughput. Measured with the common protocol of App. D.7, the DistilBERT student processes $2{,}785$ sentences/s and the $K=5$ BERT-base teacher 280 ($9.9\times$).

Summary. On SST-2 the student matches the teacher's accuracy, and its score is better than the teacher's for OOD detection and selective prediction. The variance-contraction premise of §4 does not hold on this benchmark.

F.3 Deployment-Time Experiments: Suitability Monitoring and Cascade Routing

Two deployment uses of the student's score: a label-free suitability-monitoring test on CIFAR-10-C (§F.3.1) and cascade routing based on the student's score (§F.3.2).

F.3.1 SUITABILITY MONITORING ON CIFAR-10-C

This subsection details the monitoring experiment of §5.5; the test is the one analysed in Theorem 3. The premises of Theorem 3(c) are $\sigma_1(u_S) \leq \sigma_1(u_T)$ and a preserved mean shift on the stream. We check them per scenario against the reference profile below. $\sigma_1(u_S) \leq \sigma_1(u_T)$ holds in all 75 scenarios (median ratio 0.75). The mean shift is not preserved: the student's shift is smaller than the teacher's in 66 scenarios (median ratio 0.82). The standardised shift Δ/σ_1, which determines the power of the test, is at least the teacher's in 67 scenarios; the exceptions are pixelate at every severity, defocus blur at severity 1, and brightness at severities 1–2. The premises therefore hold only in part, and the higher power is an empirical observation that Theorem 3(c) does not fully explain.

Datasets. In-distribution: CIFAR-10. Deployment stream: CIFAR-10-C (Hendrycks & Dietterich, 2019), 15 corruption types $\times$ 5 severities $=$ 75 scenarios. The standard 10,000-image CIFAR-10 test set is partitioned (deterministic, seed=0) into a 2,000-image LR training subset (Suitability Filter only), a 5,000-image reference profile R, and a 3,000-image null pool for empirical FPR estimation. CIFAR-10-C is built from the same 10,000 test images, so half of the images in the stream also appear, uncorrupted, in the reference profile; all eight monitors share this setup. Two of the fifteen types, Gaussian noise and contrast, also appear, together with a Gaussian blur, among the Tier-2 inputs used to train our student's head, each at one fixed severity (§3.4). Excluding these two types (65 scenarios) leaves the ranking of the monitors unchanged: macro power is 0.933 for ours, 0.927 for the teacher and at most 0.911 for the others.

Methods. We compare eight monitors. Each runs the one-sided mean-shift test of App. A.5 at α=0.05 on a per-sample score averaged over the deployment batch: M1 *student* (the student's score, single forward pass); M2 *teacher* (K=16 ensemble MI); M3 *Suitability Filter* (Pouget et al., 2025) (12 features extracted from raw logits, calibrated logistic regression); M4 *MSP gap* (single-model softmax); M5 *MC Dropout* (T=16 predictive entropy); M6 *EDL* (Dirichlet expected MI); M7 *DEUP* (residual uncertainty head); M8 *SWAG* (16 weight samples, predictive entropy).

Protocol. For each cell, 200 deployment batches of size n=1,000 are drawn without replacement; we report empirical power as the fraction of draws for which the test rejects H_0. Empirical FPR is estimated by the same protocol on the null pool. Non-inferiority margin m=0 throughout.

Results. Macro power across all 75 scenarios is in Table 52, per-corruption power in Table 53. Sample-complexity curves at severity 3 are in Figure 6. (i) All eight monitors have FPR at most 0.015 on the null pool, below the nominal α=0.05. (ii) The student's score has the highest macro power (0.932), above the K=16 teacher (+0.007) and the Suitability Filter (+0.022), with a single forward pass. (iii) The student's score and the teacher's both reach 80% power at the smallest deployment batch size ($n^\star$=50); all six other monitors require $\geq$ 100.

Table 52: **Label-free monitoring on CIFAR-10-C** (75 scenarios; one-sided test at α=0.05 on the batch mean of each score). $n^\star$: smallest batch size with 80% power at severity 3. Passes: forward passes per input.

Monitor	FPR ↓	Macro power ↑	$n^\star$ ↓	Passes ↓
MSP gap (single model)	0.000	0.909	100	1
EDL EU	0.000	0.903	100	1
DEUP EU	0.005	0.904	100	1
MC Dropout	0.015	0.868	500	16
SWAG	0.015	0.882	100	16
Suitability Filter (Pouget et al., 2025) (12 signals)	0.005	0.910	100	1
Teacher (K=16)	0.005	0.925	50	16
Ours	0.005	0.932	50	1

Corruptions with strong distributional signatures (the noise family, the blur family except defocus, snow, elastic, JPEG) are detected with at least 98% power by every monitor. The monitors differ on defocus blur, fog, brightness and contrast. On these four, the student's score is best on fog and contrast, ties the teacher on brightness, and is behind the Suitability Filter (0.820) and DEUP (0.819)

Table 53: Per-corruption macro-TPR ↑ (averaged over severities 1–5) on CIFAR-10-C, α=0.05, n=1,000, 200 replicates per cell. **Bold** marks the best non-teacher cell per row.

Corruption	MSP	EDL	DEUP	MC Dropout	SWAG	Suitability Filter	Teacher	Ours
gaussian_noise	1.000	1.000	1.000	1.000	1.000	1.000	1.000	**1.000**
shot_noise	1.000	1.000	1.000	1.000	1.000	1.000	1.000	**1.000**
impulse_noise	1.000	1.000	1.000	1.000	1.000	1.000	1.000	**1.000**
defocus_blur	0.806	0.792	0.819	0.491	0.772	**0.820**	0.801	0.811
glass_blur	1.000	1.000	1.000	1.000	1.000	1.000	1.000	**1.000**
motion_blur	1.000	1.000	1.000	0.980	1.000	1.000	1.000	**1.000**
zoom_blur	1.000	1.000	1.000	0.987	1.000	1.000	1.000	**1.000**
snow	1.000	1.000	1.000	1.000	0.999	1.000	1.000	**1.000**
frost	0.998	0.999	1.000	1.000	0.956	1.000	1.000	**1.000**
fog	0.727	0.692	0.745	0.717	0.593	0.714	0.754	**0.819**
brightness	0.372	0.328	0.343	0.420	0.164	0.350	0.493	**0.495**
contrast	0.798	0.799	0.787	0.440	0.748	0.814	0.820	**0.852**
elastic_transform	1.000	1.000	1.000	1.000	1.000	1.000	1.000	**1.000**
pixelate	0.932	0.939	0.871	0.988	0.993	0.953	1.000	**0.999**
jpeg_compression	1.000	1.000	1.000	1.000	1.000	1.000	1.000	**1.000**

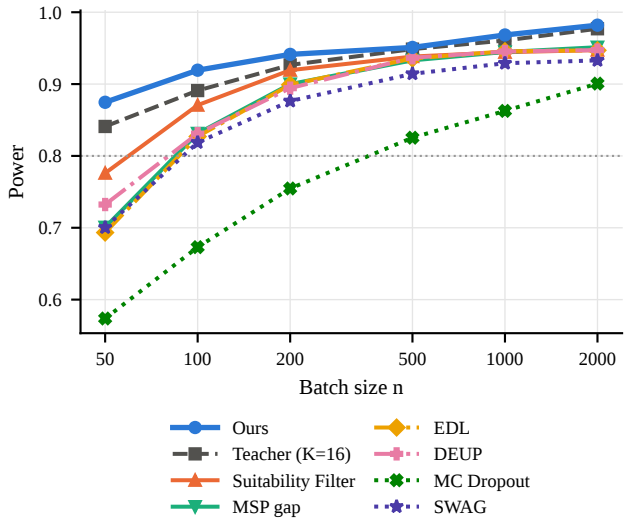

Figure 6: **Sample complexity to detect moderate shift (severity 3, averaged across all 15 CIFAR-10-C corruptions).** The dotted line marks 80% power. Our student's score and the teacher's reach 80% power at n=50 deployment samples; the six other monitors need larger batches.

on defocus blur (0.811). On every other corruption it is at least as high as every other non-teacher monitor.

F.3.2 Cascade Routing with the Student's Score

We route the $X\%$ most uncertain inputs, according to the student's score, to the K=16 teacher and accept the student's prediction otherwise.

Setup. Given a deployment input x, compute the student's score $u_S(x)$. If $u_S(x)$ is in the top-$X\%$ of scores across the deployment batch, route x to the K=16 teacher (cost $K\times$ a single forward pass); otherwise accept the student's prediction (cost $1\times$). The effective per-sample inference cost relative to a single forward pass is

$$\text{cost}(X) \;=\; 1 + X \cdot K, \tag{22}$$

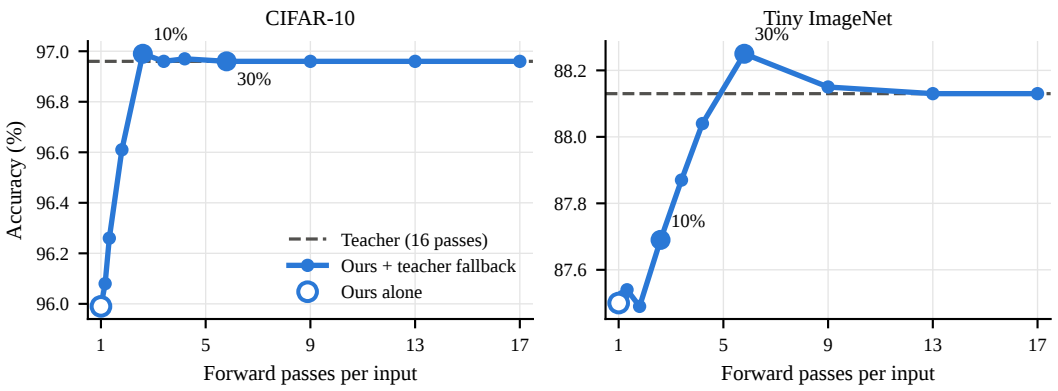

Figure 7: **Cascade routing accuracy vs. inference cost.** *Left:* CIFAR-10. *Right:* Tiny ImageNet. The fraction X of inputs with the highest scores under our student is routed to the teacher; the cost is $1 + 16X$ forward passes per input. Open circle: our student alone; labelled points: $X = 10\%$ and 30%.

since every input first passes through the student to obtain its score; this is $2.6\times$ at $X{=}10\%$ and $5.8\times$ at $X{=}30\%$ for $K{=}16$ (the standalone teacher costs $K{=}16\times$). We sweep $X \in [0\%, 100\%]$ and report accuracy and effective cost.

Table 54: Cascade routing: accuracy ↑ and effective inference cost (in parentheses, in single forward passes). **Bold** marks the best accuracy per row and the cost reduction relative to the teacher.

Dataset	Student Acc. ↑ (1×)	Teacher Acc. ↑ (16×)	Cascade@10% Acc. ↑ (2.6×)	Cascade@30% Acc. ↑ (5.8×)	Cost reduction teacher / cascade ↑
CIFAR-10	95.99%	96.96%	**96.99%**	96.96%	**6.2**× at 10%
Tiny ImageNet	87.50%	88.13%	87.69%	**88.25%**	**2.8**× at 30%

Results. On CIFAR-10, cascade@10% reaches 96.99% accuracy (teacher: 96.96%) at 2.6 forward passes per input, $6.2\times$ fewer than the teacher. The student and the teacher agree on 97.47% of inputs, and the student's score selects most of the inputs on which the teacher corrects the student.

F.4 Repeated Self-Distillation and the Scale of the Score

We build a chain $T \to S_0 \to S_1 \to \cdots \to S_R$ in which round r uses the student of round $r-1$ as its teacher, and evaluate each round on a fixed 1,000-input grid stratified by region (ID, near-OOD, far-OOD).

MNIST (3 rounds, 5 seeds). We report the head's output before any output non-linearity. From round 1 its mean on far-OOD inputs is negative ($+0.83$ at round 0, -0.56 at round 1, -1.38 at round 3), while its Pearson correlation with the previous round's score on the MNIST test set rises from 0.93 at round 1 to 0.97 at round 3, and its across-seed spread on far-OOD grows from 0 to 0.49 (Table 55). On the full grid the picture differs: because the OOD scores change sign, round 1 is negatively correlated with round 0 (Pearson -0.84; far-OOD alone -0.58), and the correlation recovers only by round 3 (0.98). High fidelity on ID inputs therefore does not preserve the absolute scale of the score on OOD inputs.

Constraining the output. We also ran rounds 1–3 (3 seeds each) with a Softplus output. All negative values disappear, but the scores collapse towards zero and the far-minus-ID gap changes sign in every constrained run. A Softplus output therefore does not solve the problem. A log-parameterised head with a scale-aware initialisation is left for future work.

Table 55: Repeated self-distillation on MNIST, 5 seeds: mean head output (before any output nonlinearity) per region, with within-grid standard deviation ± across-seed standard deviation. From round 1 the far-OOD mean is negative.

Round	Region	Mean head output (signed)	within-grid std ± across-seed std
0	ID	+0.0041	0.049 ± 0.000
	near	+0.3148	0.131 ± 0.000
	far	+0.8316	0.107 ± 0.000
1	ID	+0.0046	0.075 ± 0.014
	near	−0.0238	0.185 ± 0.059
	far	−0.5642	0.180 ± 0.185
2	ID	+0.0034	0.094 ± 0.015
	near	−0.0506	0.232 ± 0.077
	far	−0.5896	0.162 ± 0.200
3	ID	+0.0055	0.107 ± 0.016
	near	−0.1831	0.378 ± 0.076
	far	−1.3759	0.517 ± 0.488

CIFAR-10 (5 rounds, 2 seeds). The within-grid standard deviation drops sharply at round 1 (far-OOD $0.084 \to 0.019$) and stays below round 0 through round 5, but the far-OOD mean also collapses at round 1 ($0.444 \to 0.175$) and never recovers, and the ID mean drifts upward by about 0.013 per round (Table 56). The student's accuracy after five rounds is below round 0.

Table 56: Repeated self-distillation on CIFAR-10: mean score and within-grid standard deviation per round, averaged over 2 seeds. **Bold**: smallest within-grid standard deviation per region.

	R0		R1		R2		R3		R4		R5	
Region	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
ID (CIFAR-10)	0.073	0.098	0.079	**0.056**	0.098	0.064	0.114	0.072	0.129	0.081	0.139	0.086
Near-OOD (CIFAR-100)	0.318	0.134	0.169	**0.040**	0.196	**0.040**	0.221	0.045	0.242	0.051	0.257	0.056
Far-OOD (SVHN)	0.444	0.084	0.175	0.019	0.205	**0.017**	0.232	0.019	0.256	0.032	0.275	0.027

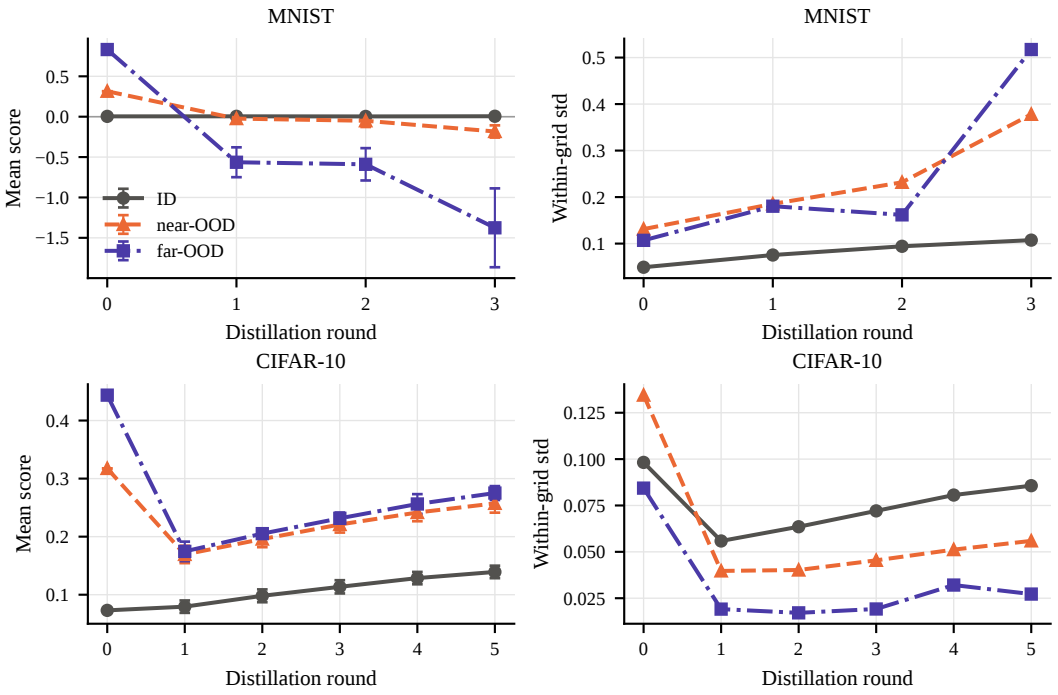

Figure 8: **Repeated self-distillation.** *Top:* MNIST (3 rounds, 5 seeds). *Bottom:* CIFAR-10 (5 rounds, 2 seeds). *Left:* per-round mean score by region (MNIST: head output before any output non-linearity), with the across-seed standard deviation as error bars (zero at round 0, where all seeds share one teacher). *Right:* within-grid standard deviation (Tables 55 and 56). ID = MNIST/CIFAR-10 test, near-OOD = FashionMNIST/CIFAR-100, far-OOD = SVHN.