

Trust the Uncertain Teacher: Distilling Dark Knowledge via Calibrated Uncertainty

Jeonghyun Kim^{1,2}, Richeng Xuan³, Hyung Wook Noh^{2,4},
Soo Kyung Kim^{1†}, Hyunsoo Cho^{1†*}

¹Ewha Womans University, ² Electronics and Telecommunications Research Institute,
³ Beijing Academy of Artificial Intelligence, ⁴ University of Science and Technology
{jhyunye, chohyunsoo}@ewha.ac.kr

Abstract

The core of knowledge distillation lies in transferring the teacher’s rich ‘dark knowledge’—subtle probabilistic patterns that reveal how classes are related and the distribution of uncertainties. While this idea is well established, teachers trained with conventional cross-entropy often fail to preserve such signals. Their distributions appear decisive but are in fact brittle, offering little beyond the hard label. This overconfidence is especially problematic in high-cardinality tasks, where subtle distinctions among many plausible classes are crucial for guiding compact students. To address this limitation, we revisit distillation from a distributional perspective and propose Calibrated Uncertainty Distillation (CUD), a framework that makes dark knowledge more faithfully accessible. Instead of uncritically inheriting the teacher’s overconfidence, CUD encourages the teacher to expose informative uncertainty and guides the student to learn from calibrated targets instead of overly sharpened predictions. By shaping the teacher’s predictive distribution before transfer, our approach balances accuracy and calibration, allowing students to benefit from both confident signals on easy cases and structured uncertainty on hard ones.

1 Introduction

LLMs such as GPT-4 (Achiam et al., 2023), PaLM (Chowdhery et al., 2023), LLaMA (Touvron et al., 2023), and Gemini (Team et al., 2023) deliver state-of-the-art results across reasoning, dialogue, and code generation, yet they impose heavy latency, memory, and energy costs at inference time. Among compression techniques including pruning and quantization, knowledge distillation (KD) (Hinton et al., 2015) remains uniquely attractive because it transfers a teacher’s *predictive distribution*,

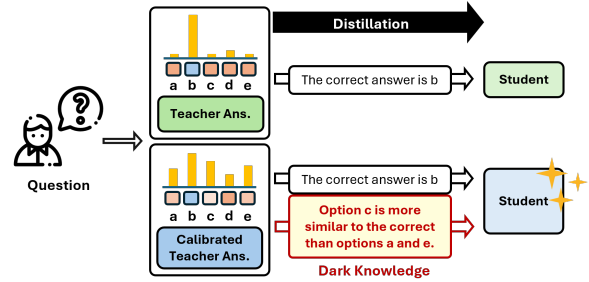


Figure 1: Calibrated teacher distributions expose informative uncertainty, facilitating more effective transfer of dark knowledge during distillation.

not just argmax labels, thereby capturing inter-class structure, inductive biases, and task-dependent uncertainty that smaller models would not infer from hard labels alone (Huang et al., 2022; Moslemi et al., 2024). Building on the perspective that calibrated teacher distributions can more faithfully reflect relational structure among plausible classes (Dong et al., 2024), our work aims to preserve and selectively transfer this structure to support students that generalize more reliably.

However, conventional supervised training often yields an over-confident teacher (Wei et al., 2022; El Baida et al., 2025) whose probabilities collapse onto a single class, suppressing precisely the soft structure that makes KD effective. Post-hoc temperature scaling, which adjusts prediction sharpness with a global scalar after training, only rescales logits and cannot recover information that was never represented or was erased by overfitting (Guo et al., 2017). Regularizers such as label smoothing and confidence penalties may reduce calibration error, yet they can also blur informative class-wise geometry that students should imitate (Müller et al., 2019; Pereyra et al., 2017). Much of the KD literature compensates on the student side while keeping the teacher fixed, for example via KL matching with a fixed temperature, intermediate feature hints, or self-distillation (Zhang

*Corresponding author

[†]Department of AI, Institute for Multiscale Matter and Systems

et al., 2019; Mobahi et al., 2020). These strategies are valuable, but they neither prevent distribution collapse in the teacher nor selectively filter teacher errors during transfer, leaving performance on borderline and long-tail examples especially fragile.

We take a distributional perspective on KD and argue that two properties are jointly required for strong downstream students. First, the *teacher* should retain informative uncertainty, allocating non-trivial probability mass to semantically related alternatives where the task is intrinsically ambiguous. Second, the *transfer* should be calibrated so that the student imitates informative relations while attenuating misleading teacher peaks that arise from over-confidence or spurious correlations. Intuitively, KD works best when the teacher behaves like a well-calibrated Bayesian posterior on easy and hard inputs alike, and when the student is encouraged to inherit that calibrated structure rather than the teacher’s certainty alone. We define these characteristics in problem formulation section.

This view aligns with the empirical regularities observed across the modalities. In fine-grained classification and language understanding, soft targets consistently improve sample efficiency by encoding *relative* similarities between classes that hard labels cannot convey. When the data distribution shifts, students distilled from calibrated teachers signal their own mistakes more reliably and maintain more stable confidence on near-OOD inputs. This stability leads to improved selective prediction and more favorable accuracy–rejection trade-offs (Guo et al., 2017; Mishra et al., 2023). In resource-constrained deployments, such as on-device assistants (Chi et al., 2025) and retrieval-augmented pipelines where latency budgets are tight (Jiang et al., 2025), students that inherit calibrated uncertainty achieve competitive accuracy while maintaining predictable confidence estimates, a property that downstream systems can exploit for routing, abstention, and cost-aware ensembling.

Our study operationalizes this distributional stance without relying on architectural tricks or task-specific heuristics. We focus on *why* preserving and selectively transferring dark knowledge is necessary and *when* it pays off, rather than on implementation details. Empirically, we show that students trained under this lens achieve higher accuracy at fixed compute and improved robustness on ambiguous and long-tail slices.

The results support the thesis that effective KD

should prioritize calibrated structure over sharpened certainty, allowing the student to benefit from richer relational signals while avoiding the teacher’s overconfident mistakes.

2 Problem Formulation

2.1 Preliminaries and Limitations of Conventional KD

Let $\mathcal{X}$ be the input space and $\mathcal{Y} = \{1, \dots, C\}$ the label set. A teacher model induces a predictive distribution $p_T(\cdot | x) \in \Delta^{C-1}$ for $x \in \mathcal{X}$, and a student model with parameters θ_S induces $p_S(\cdot | x; \theta_S) \in \Delta^{C-1}$. Conventional knowledge distillation (KD) trains the student to match the teacher (optionally combined with ground-truth supervision):

$$\min_{\theta_S} \mathbb{E}_{x \sim \mathcal{D}} [\mathcal{L}_{\text{KD}}(p_S(\cdot | x; \theta_S), \tilde{p}_T(\cdot | x))] + \lambda \mathbb{E}_{(x,y) \sim \mathcal{D}} [\mathcal{L}_{\text{CE}}(p_S(\cdot | x), y)], \quad (1)$$

where $\tilde{p}_T(\cdot | x)$ denotes the *calibrated* teacher target (usually just a temperature scaled version of p_T), and $\mathcal{D}$ is the labeled set. In practice, however, maximum likelihood training often produces an over confident p_T , so that even such temperature scaled targets can remain over sharpened, erasing relational “dark knowledge” and amplifying erroneous peaks that are then propagated to p_S .

2.2 Guiding Principles for Calibrated Distillation

We posit that KD can be simultaneously *robust* and *informative* when the following conditions hold.

C1: Difficulty-aware uncertainty. For ambiguous or challenging inputs, the teacher should preserve uncertainty corresponding to task difficulty rather than collapsing probability mass onto a single class. Concretely, the teacher should retain non-trivial probability on semantically related alternatives, thereby encoding relational cues that are essential for generalization.

C2: Calibrated, selective imitation. The student should not blindly replicate the teacher’s distribution. Instead, it should selectively imitate the teacher’s informative relational structure while suppressing spurious over-confident peaks arising from teacher errors. This preserves the benefits of the teacher’s nuanced knowledge without propagating its mistakes.

2.3 Calibration as a Constraint-based Reformulation

Motivated by the above requirements, we recast the knowledge distillation problem as a teacher calibration problem: we seek a calibration operator $\mathcal{C}$ such that, for each input x , the calibrated target $\tilde{p}_T(\cdot | x) = \mathcal{C}(p_T(\cdot | x), x)$ is obtained by projecting the raw teacher distribution onto the set of distributions that satisfy the following constraints R1 and R2:

R1: Uncertainty with semantic support (for difficulty-aware uncertainty). Maintain difficulty-appropriate concentration and guarantee mass over a stable neighborhood of semantically related classes:

$$\begin{aligned} h_{\min}(x) &\leq H(\tilde{p}_T(\cdot | x)) \leq h_{\max}(x), \\ \sum_{k \in \mathcal{N}(y; x)} \tilde{p}_T(k | x) &\geq \alpha(x). \end{aligned}$$

Here $H(\cdot)$ is an uncertainty measure (e.g., entropy), and $\mathcal{N}(y; x)$ is a semantics-preserving neighborhood of the true label y (e.g., defined via taxonomies, embeddings, or empirical confusion).

R2: Wrong-mass budget (for selective imitation). Bound probability assigned to evidently incompatible classes:

$$\sum_{k \in \mathcal{W}(x, y)} \tilde{p}_T(k | x) \leq \varepsilon(x),$$

where $\mathcal{W}(x, y)$ denotes classes that carry overconfident wrong mass for (x, y) (e.g., via rules, weak labels, or high-confidence constraints).

2.4 Projection to Calibrated Targets

Among all distributions satisfying these constraints, we select the one closest to the original teacher in a structure-preserving sense:

$$\begin{aligned} \tilde{p}_T &\in \arg \min_{q \in \Delta^{C-1}} \text{Dist}(q, p_T(\cdot | x)) \\ \text{s.t. } &\text{R1, R2,} \end{aligned} \quad (2)$$

where Dist is a structure-sensitive divergence (e.g., symmetrized KL or Wasserstein). Coupled with the training objective in (1), this projection emphasizes calibrated pairwise ratios and ordinal relations rather than brittle peak sharpening. We prove in Appendix A that the optimization problem in (2) admits a unique global minimizer without spurious local minima.

3 Calibrated Uncertainty Distillation

While Eq. (2) establishes the theoretical ideal for target calibration, solving this constrained optimization problem directly for every training sample is computationally intractable. To bridge this gap, we propose Calibrated Uncertainty Distillation (CUD) designed to *relax* the reformulated constraints (R1, R2) in a practical and modular way. Specifically, our practical modules, DUS and W-Clip, efficiently approximate the optimal solution: (i) **DUS**, which reshapes the teacher’s confidence to counter overconfidence while retaining sharp predictions on easy cases, and (ii) **W-Clip**, which adjusts target distributions to suppress spurious peaks without disturbing relational structure. Together, these modules make the constraints operational and yield calibrated soft targets that improve knowledge transfer.

3.1 Difficulty-aware Uncertainty Shaping

DUS modifies the teacher distribution to provide uncertainty where it is most useful. The key idea is to *increase entropy selectively*: ambiguous or misclassified inputs are encouraged to carry higher uncertainty, while confident and correct predictions remain sharp. Concretely, we aim to raise entropy only when the teacher is uncertain or wrong—satisfying the lower bound $h_{\min}(x)$ —and avoid unnecessary blurring on confident, correct cases to respect the upper bound $h_{\max}(x)$ and maintain class geometry.

Let $p_T(\cdot | x)$ be the teacher distribution and $p_T^y := p_T(y | x)$ for ground-truth y . We fine-tune the teacher with a **focal-entropy** objective that down-weights trivial cases and rewards non-degenerate uncertainty on hard or mispredicted inputs:

$$\begin{aligned} \mathcal{L}_{\text{teacher}} &= \lambda_{\text{CE}}(-\log p_T^y) \\ &\quad + \lambda_{\text{F}}(-\alpha_y(1 - p_T^y)^\gamma \log p_T^y) \\ &\quad - \lambda_{\text{H}} w(x) H(p_T(\cdot | x)), \end{aligned} \quad (3)$$

where $\alpha_y \in (0, 1]$, $\gamma \geq 0$, and $H(\cdot)$ denotes entropy. The difficulty gate

$$w(x) = \mathbb{1}\{p_T^y < \tau\} + \rho(1 - p_T^y)^\beta \quad (4)$$

activates entropy shaping only when the teacher is wrong or low-confidence. The focal factor $(1 - p_T^y)^\gamma$ reduces gradients on trivially correct examples ($p_T^y \approx 1$) and emphasizes hard/mispredicted ones (p_T^y small); combined with the gated entropy reward $w(x) H(\cdot)$, this selectively increases

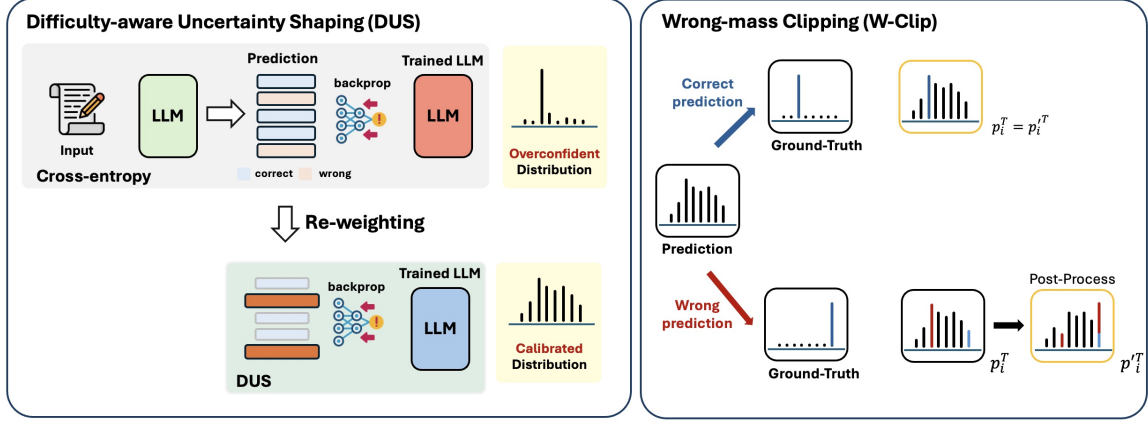


Figure 2: Illustration of the main framework. Left: Difficulty-aware Uncertainty Shaping (DUS) adjusts weights based on the correctness of the model’s predictions, reducing overconfidence and encouraging a more calibrated probability distribution. Right: Wrong-mass Clipping (W-Clip) modifies the probability assigned to incorrect classes when the model makes a wrong prediction, providing a more stable and reliable distribution for the student.

entropy where it improves distillation (meeting $H(\tilde{p}_T) \geq h_{\min}(x)$) while leaving easy cases sharp (implicitly respecting $h_{\max}(x)$) without resorting to global label smoothing.

3.2 W-Clip: Wrong-mass Clipping

W-Clip calibrates the transfer by reallocating probability mass from over-confidently wrong predictions. Instead of reshaping the full distribution, W-Clip applies a budgeted correction that suppresses an overconfident top-1 error when it outranks the ground truth. Formally, we approximate the wrong-mass constraint $\sum_{k \in \mathcal{W}(x,y)} \tilde{p}_T(k | x) \leq \varepsilon(x)$ by suppressing only the wrong top-1 class under a small, controllable budget.

Let $k^* = \arg \max_k p_T(k | x)$ and suppose $k^* \neq y$. We define two gated amounts

$$\delta_{\text{budget}}(x) \triangleq \eta p_T(k^* | x), \quad (5)$$

$$\delta_{\text{margin}}(x) \triangleq m (p_T(k^* | x) - p_T(y | x)), \quad (6)$$

where $\eta \in (0, 1)$ caps the total removable mass (the *wrong-mass budget*) and $m \in (0, 1]$ scales suppression with the teacher’s over-confidence margin. We then set

$$\delta(x) = \min\{\delta_{\text{budget}}(x), \delta_{\text{margin}}(x)\}, \quad (7)$$

$$\begin{aligned} \tilde{p}_T(y | x) &= p_T(y | x) + \delta(x), \\ \tilde{p}_T(k^* | x) &= p_T(k^* | x) - \delta(x), \end{aligned} \quad (8)$$

$$\tilde{p}_T(k | x) = p_T(k | x), \quad \text{for } k \notin \{y, k^*\}. \quad (9)$$

If $k^* = y$, we simply take $\tilde{p}_T = p_T$. The budget term (5) directly implements the constraint that

Margin $M(x)$	Banking77	CLINC150	MASSIVE
< 0.05	0.056	0.055	0.066
$0.05-0.15$	0.256	0.289	0.246
$0.15-0.30$	0.428	0.374	0.392
$0.30-0.50$	0.493	0.480	0.478
> 0.50	0.500	0.500	0.500

Table 1: **Relative clipping ratio stratified by over-confidence margin.** The reported value is $\delta(x)/p_T(k^* | x)$, with $M(x) = p_T(k^* | x) - p_T(y | x)$. Near-neighbour errors are barely touched; the budget $\eta = 0.5$ is reached only for confidently wrong predictions.

only a limited fraction of wrong-class mass can remain, while the margin term (6) focuses corrections on strongly over-confident mistakes. The amount removed therefore scales with the over-confidence margin $p_T(k^* | x) - p_T(y | x)$, up to the cap $\eta p_T(k^* | x)$: when the wrong top-1 is a semantic neighbour of the ground truth it receives similar probability to it, so little mass is removed. Table 1 shows that over 93% of the mass is preserved at margins below 0.05, while η is reached only when the teacher places a wrong class far above the ground truth. Keeping all non-involved classes unchanged preserves pairwise odds (dark knowledge), so the student inherits informative relations without propagating the teacher’s spurious peak.

3.3 Student Distillation with Calibrated Targets

Given the calibrated targets $\tilde{p}_T(\cdot | x)$, the student minimizes a tempered KL (plus optional CE on

labeled data):

$$\min_{\theta_S} \mathbb{E}_x \left[\lambda_{\text{KD}} \tau^2 \text{KL}(\tilde{p}_T^{(\tau)}(\cdot | x) \| p_S^{(\tau)}(\cdot | x; \theta_S)) \right] + \lambda_{\text{CE}} \mathbb{E}_{(x,y)} \left[-\log p_S(y | x) \right], \quad (10)$$

where $q^{(\tau)} = \text{softmax}(\log q / \tau)$ and $\tau > 0$. This transfers the calibrated structure rather than brittle peaks; unlabeled examples contribute only the KD term, supporting semi-supervised settings. In this case W-Clip is naturally bypassed—since detecting a wrong top-1 requires a ground-truth label—and the student learns directly from the DUS-shaped teacher distribution.

3.4 Operational Realization of Constraints

In Section 2, we formulated target calibration as a constrained projection problem. Here, we connect our practical modules to this framework by relaxing the abstract sets $N(y; x)$ and $W(x, y)$ for the constraints R1 and R2.

Operationalizing Constraint Sets. To render the projection tractable, we instantiate these sets based on the principal modes of teacher failure:

1. Operational Wrong-Mass Set (W_{op}): We define $W_{op}(x, y) = \{k^*\}$ strictly when the top-1 prediction k^* outranks the ground truth ($k^* \neq y$), and $\emptyset$ otherwise. This definition targets the dominant source of over-confident wrong mass identified in R2.

2. Operational Neighborhood Set (N_{op}): Rather than explicitly defining a semantic neighborhood, we operationalize R1 implicitly. Maximizing entropy via DUS widens the distribution, satisfying the neighborhood mass requirement without pre-defined similarity graph.

Approximating the Optimal Projection. As proven in Appendix A, the optimal solution q^* to the constrained projection is an *exponential tilting* of the teacher distribution. However, computing the partition function for this tilt is computationally expensive. We demonstrate in Appendix A.3 that **W-Clip** constitutes a *first-order Taylor approximation* of this exponential tilt around $\nu = 0$. Specifically, the linear subtraction in W-Clip recovers the analytical solution ($e^{-\nu} \approx 1 - \nu$) for conservative corrections. Similarly, **DUS** acts as a regularizer that pre-conditions the teacher to satisfy the entropy bounds of R1. Thus, CUD replaces the intractable iterative projection with efficient $O(1)$ arithmetic operations grounded in information geometry.

4 Experimental Results

4.1 Training Setup

Models & Protocol. We structure evaluation in two parts. **(Teacher-side)** We first train a standard *teacher*—bert-base-uncased (Devlin et al., 2019) (12 layers, 768 hidden)—and compare its performance against our distillation-friendly teacher (DUS) under the same data splits. **(Student-side)** We then distill from the trained teacher and compare students across KD methods, using two capacity variants to probe different architectural gaps: **Mini** (4 layers, 256 hidden), which changes both depth and width so feature-based distillation is inapplicable, and **Small** (6 layers, 768 hidden), which preserves hidden size and thus admits feature-KD baselines; our method uses calibrated response targets in both cases. All results are reported as the mean and variance computed over three independent runs.

Hyperparameters & Rationale. While our method introduces several hyperparameters, most show low sensitivity; we tuned them once on the Banking dataset and used the same global values for all tasks, avoiding per-dataset hyperparameter engineering (see Appendix B). The key distillation parameters include the entropy reward $\lambda_H = 0.1$, the focal parameters $\gamma = 10$ and $\alpha_y = 1$, and the W-Clip budget $\eta = 0.5$ and margin scaling $m = 0.7$. For the difficulty gate in Eq. (4), we set the confidence threshold $\tau = 0.7$, the continuous scaling factor $\rho = 1.0$, and the difficulty exponent $\beta = 1.0$, and together these values behave such that smaller settings recover a standard distillation baseline, while larger settings amplify the inductive bias of our approach. In addition, several minor scalar ratios λ_{KD} and λ_{CE} balance the cross-entropy and KD objectives. All experiments use AdamW with batch size 32, maximum sequence length 128, weight decay 0.01, gradient clipping at 1.0, and 10% linear warmup with cosine decay.

Benchmarks & Evaluation. We evaluate our method on five classification datasets spanning a wide range of label granularities, from 4 to 150 classes: Banking77 (77 intents) (Casanueva et al., 2020), CLINC150 (150 intents) (Larson et al., 2019), MASSIVE (60 intents) (FitzGerald et al., 2023), TREC (6 question types) (Li and Roth, 2002), and AG News (4 topic categories) (Zhang et al., 2015). Detailed dataset descriptions are provided in the Appendix C. Additional results on

binary classification settings are also included in the Appendix D for completeness.

Baselines. We compare against representative KD methods, summarized by their characteristics:

1. **LKD (Logit-KD)** (Hinton et al., 2015): Response-based KL matching on teacher soft logits (with temperature). Simple and task-agnostic but ignores intermediate structure.
2. **PKD (Patient KD)** (Sun et al., 2019): Layer-wise alignment of hidden representations. *Exploits intermediate features and depth supervision*; effective when student width/layers can be mapped to the teacher, less suited to severe architectural mismatch.
3. **TinyBERT** (Jiao et al., 2020): Distills both self-attention and hidden states, originally with two-stage training and data augmentation. *Achieves high compression quality via rich feature-level objectives*; in our setup, we disable augmentation for a fair task-specific comparison.
4. **CKD (Contextual KD)** (Park et al., 2021): Transfers *relational knowledge* by modeling two forms of contextual structure: *Word Relation* between tokens, and *Layer Transforming Relation* across layers. Architecture-agnostic and effective for heterogeneous teacher–student pairs.
5. **MGSKD (Multi-Granularity Structural KD)** (Liu et al., 2022): Distills *structural relations* over *multiple semantic granularities* (tokens, spans, samples), formulated via pair-wise interactions and triplet geometric angles. Preserves richer linguistic structure than mono-granularity KD, at the cost of higher relational computation.
6. **AD-KD** (Wu et al., 2023): Response-based distillation with confidence/uncertainty-aware reweighting or calibration. *Mitigates propagation of over-confident teacher errors* while keeping architecture independence and low overhead.

4.2 Main results

Table 2 reports both teacher side and student side performance. Across both student capacities, CUD outperforms prior baselines on average.

Effect of student capacity. For the larger student (6L 768D), several baselines already achieve accuracy comparable to the teacher (teacher average 94.32). In this regime, CUD still attains the best average student performance (94.33), improving over LKD by about +1.17 point on average (94.33 versus 93.16). For the smaller student (4L 256D), the dimensionality gap to the teacher makes representation–level matching difficult or even infeasible, so performance is largely determined by the quality of the response targets. In this regime, CUD substantially outperforms other baselines by a larger margin.

Effect of cardinality. CUD yields much larger gains, especially on tasks with many classes such as Banking77, CLINC150, MASSIVE, where it improves over LKD by +10.43, +5.14, and +5.22 points respectively. This pattern matches our intuition: on tasks with many classes, the calibrated targets keep meaningful probability on several plausible labels while suppressing the dominant wrong peak, so the student can imitate a richer set of alternatives. In contrast, when the number of classes is small, for example (AGNews, TREC), the teacher distribution is already almost deterministic and offers little additional structure to recover. For extremely small numbers of classes (binary), this effect becomes stronger; suppressing a wrong peak acts like mild logit re-tempering and adds little information beyond standard smoothing or temperature scaling. Detailed analysis is provided in the Appendix D.

Comparison with teacher-side methods. Because our contribution lies in how the teacher is prepared rather than in the student objective, we compare against other interventions on the teacher distribution: post-hoc rescaling (temperature and vector scaling), regularized training (label smoothing, Mixup), and the MSE teacher of Hamidi et al. (2024), all under an identical teacher–student architecture. Table 3 shows a clear ordering. A single global temperature or a fixed smoothing coefficient gains almost nothing over the standard teacher, as the same transformation applied to every sample leaves the teacher’s confidence structure intact; per-class parameters and teacher retraining gain several points. CUD is best throughout, and what separates it from these stronger baselines is that its adjustment is chosen per sample rather than fixed in advance.

Dataset	Banking77	CLINC150	TREC	AGNews	Massive	Avg
FT Teacher	93.70	96.35	97.40	94.69	89.47	94.32
Student (6L-768D)						
TinyBERT [†]	89.47 ± 0.50	88.43 ± 0.06	96.13 ± 0.12	94.18 ± 0.15	88.84 ± 0.40	91.41
MGSKD [†]	91.58 ± 0.04	95.42 ± 0.05	87.60 ± 0.20	94.64 ± 0.15	88.56 ± 0.32	91.56
PKD	89.87 ± 0.15	95.09 ± 0.17	96.40 ± 0.20	94.54 ± 0.24	88.17 ± 0.20	92.81
LKD	93.03 ± 0.11	93.51 ± 0.44	96.07 ± 0.23	94.23 ± 0.29	88.94 ± 0.21	93.16
AD-KD	92.66 ± 0.45	95.42 ± 0.37	96.00 ± 0.35	94.57 ± 0.31	89.30 ± 0.28	93.59
CKD	<u>93.54</u> ± 0.24	<u>95.93</u> ± 0.17	97.53 ± 0.46	<u>94.72</u> ± 0.15	<u>89.30</u> ± 0.25	94.20
CUD (Ours)	93.83 ± 0.06	96.07 ± 0.14	<u>97.40</u> ± 0.35	94.91 ± 0.12	89.44 ± 0.06	94.33
Student (4L-256D)						
LKD	80.79 ± 0.44	88.31 ± 0.05	<u>95.80</u> ± 0.00	94.28 ± 0.04	82.00 ± 0.04	88.24
AD-KD	82.22 ± 0.57	91.35 ± 0.08	95.06 ± 0.12	94.03 ± 0.05	84.17 ± 0.47	89.37
PKD	86.60 ± 0.39	93.27 ± 0.15	<u>95.80</u> ± 0.53	94.37 ± 0.13	87.23 ± 0.07	91.45
CKD	<u>89.86</u> ± 0.19	91.70 ± 0.15	95.40 ± 0.35	<u>94.24</u> ± 0.09	86.60 ± 0.22	91.56
CUD (Ours)	91.22 ± 0.02	93.45 ± 0.05	96.47 ± 0.23	94.11 ± 0.27	<u>87.22</u> ± 0.07	92.49

[†] Methods are trained with distillation already at the pretraining stage; we fine-tune publicly available checkpoints.

Table 2: **Main experimental results.** Accuracy on five classification benchmarks for two student sizes (4L-256D and 6L-768D) distilled from a fine-tuned teacher. The best performance is highlighted in **bold**, and the second-best is underlined. Our proposed CUD attains the best average accuracy for both student sizes, with especially pronounced gains on high-cardinality tasks and for the smaller student, highlighting its effectiveness under challenging transfer regimes.

Method	Banking77	CLINC150	MASSIVE	Avg
LKD	80.79	88.31	82.00	83.70
Temp. scaling	80.61	88.35	82.04	83.67
Label smoothing	81.29	86.88	82.19	83.45
Vector scaling	87.21	93.33	86.97	89.17
Mixup	87.91	92.87	87.06	89.28
MSE teacher	88.21	93.20	86.67	89.36
CUD (Ours)	91.22	93.45	87.22	90.63

Table 3: **Comparison with teacher-side methods.** All rows distil into the same 4L-256D student and differ only in how the teacher’s distribution is produced; LKD uses the standard fine-tuned teacher. Only CUD varies the adjustment with each sample’s difficulty.

4.3 Uncertainty and OOD Robustness

Our approach shapes predictive uncertainty for transfer, which also provides more reliable confidence estimates and stronger performance on Out-of-Distribution (OOD) detection.

OOD Detection We follow a standard OOD detection protocol: the teacher and student are trained on BANKING77 (in-distribution - IND), and we treat CLINC150, MASSIVE, and other intent datasets with different domains and label spaces as OOD. We compute AUROC, FPR95, and FPR90 to quantify how well IND and OOD samples are separated. As shown in Table 4, CUD achieves the highest AUROC and the lowest FPR across all OOD benchmarks. The gains are partic-

ularly large on heterogeneous OOD datasets such as CLINC150 and MASSIVE, indicating that the uncertainty shaped by DUS transfers to stronger OOD rejection.

Uncertainty Estimation To evaluate how well each model estimates uncertainty, we compute Expected Calibration Error (ECE) and Brier score on the test set of each dataset. CUD achieves markedly lower calibration errors on wrong predictions (while overall ECE is higher due to underconfident shift) and superior correctness AUROC (0.92 vs. 0.83 for LKD), indicating that the improvement lies in how predictions are ranked rather than in the absolute scale of confidence. Detailed analysis, including ECE/Brier scores (Table 10 and Table 12) and reliability curves (Figure 3), is provided in Appendix E.

5 Analysis

5.1 Ablation Study

Table 5 reports ablations of our main modules (DUS and W-Clip) in the 4L-256D student. While both components improve over LKD on average, and the gains grow with label cardinality as noted in the main results, the detailed breakdown further reveals when and why they matter.

Closing the teacher-student gap. Relative to the fine-tuned teacher, LKD leaves substantial headroom: -12.91 points on Banking77 (80.79 vs.

Task	Method	AUROC	FPR95	FPR90
CLINC150	TinyBERT	79.06	70.04	54.55
	CKD	92.21	38.97	22.53
	AD-KD	93.91	30.08	13.31
	CUD (Ours)	94.52	24.24	12.77
TREC	TinyBERT	82.34	61.80	42.60
	CKD	95.18	31.40	15.00
	AD-KD	97.30	19.20	7.20
	CUD (Ours)	97.45	16.60	8.60
AGNews	TinyBERT	85.28	58.47	41.93
	CKD	95.66	27.50	11.86
	AD-KD	98.41	8.98	2.02
	CUD (Ours)	99.07	4.53	1.56
MASSIVE	TinyBERT	80.39	68.63	53.17
	CKD	93.74	37.45	19.37
	AD-KD	95.80	29.59	10.99
	CUD (Ours)	96.79	20.12	8.90

Table 4: **Cross-dataset OOD detection performance across student KD methods.** All teacher and student models are trained on Banking77; each row reports AUROC, FPR95, and FPR90 based on the student’s maximum softmax probability.

Dataset	LKD	LKD+DUS	LKD+W-Clip	CUD(Ours)
Banking77	80.79	90.62	89.25	91.22
CLINC150	88.31	93.24	93.02	93.45
TREC	95.80	96.40	95.40	96.47
AGNews	94.28	94.17	94.28	94.11
MASSIVE	82.00	86.76	86.37	87.22
Avg	88.24	92.24	91.66	92.49
Avg Δ	—	+4.00	+3.42	+4.25

Table 5: **Ablation of our model on multiclass benchmarks.** Performance of a 4-layer, 256-dim student under different distillation variants. Both DUS and W-Clip improve upon LKD, and their combination (CUD) yields the strongest gains.

93.70), -8.04 on CLINC150 (88.31 vs. 96.35), and -7.47 on MASSIVE (82.00 vs. 89.47). CUD closes 81%, 64%, and 70% of these gaps respectively, demonstrating that calibrated uncertainty specifically helps students recover performance where naive KD collapses.

Relative impact of modules. DUS alone outperforms W-Clip in average accuracy (92.24 vs. 91.66), showing that reshaping the teacher distribution is more fundamental than just reweighting samples. W-Clip is beneficial in high-class regimes (e.g., $+8.46$ on Banking77 and $+4.37$ on MASSIVE vs. LKD), but can over-emphasize noise in easier tasks. The full CUD system balances both: projection (R2) prevents over-amplification

Dataset	Accuracy	ECE _{wrong}	Brier _{wrong}
Banking77	93.70 $\rightarrow$ 92.43	0.817 $\rightarrow$ 0.546	1.654 $\rightarrow$ 1.173
CLINC150	96.35 $\rightarrow$ 94.93	0.790 $\rightarrow$ 0.619	1.601 $\rightarrow$ 1.275
TREC	97.40 $\rightarrow$ 97.00	0.933 $\rightarrow$ 0.924	1.790 $\rightarrow$ 1.751
AGNews	94.69 $\rightarrow$ 94.72	0.801 $\rightarrow$ 0.504	1.406 $\rightarrow$ 0.743
MASSIVE	89.47 $\rightarrow$ 89.47	0.813 $\rightarrow$ 0.623	1.641 $\rightarrow$ 1.269

Table 6: **Effect of DUS on the teacher.** Each cell reports the standard fine-tuned teacher followed by the DUS teacher. Accuracy changes by at most 1.4 points, while calibration error on the teacher’s own mistakes drops substantially on four of five datasets. TREC is the exception: at 97.40 accuracy, errors are rare and the distribution is already near-deterministic.

of wrong peaks while clipped weighting emphasizes informative errors, producing the most stable and highest average (92.49, $\Delta+4.25$).

Teacher calibration. In our setup, the DUS-trained teacher attains slightly lower accuracy than a standard fine-tuned teacher, yet students distilled from it outperform those distilled from the standard teacher. This holds not only for our CUD objective but also for other KD baselines when we simply replace their original teacher with ours (teacher side ablations in Tab. 14).

Table 6 shows what changes. Accuracy moves by at most 1.4 points and is essentially unchanged on MASSIVE and AGNews, while calibration error computed over the teacher’s own mistakes falls substantially on four of five datasets. The teacher does not become more accurate; it becomes less confident, and most sharply so where it is wrong, which is the signal a student can use.

Behavior on binary classification tasks. CUD is most effective when the teacher distribution contains rich multi-class structure with informative relative probabilities. In binary or near-binary tasks, however, the teacher’s output collapses to a one-dimensional form $(p, 1 - p)$, inherently limiting the amount of dark knowledge available for redistribution. Empirical results on QNLI and SST-2 (Appendix D) confirm this analysis: compared to high-cardinality settings, the gains from CUD are substantially smaller and sometimes comparable to LKD.

5.2 Transfer to Large Language Models

To examine whether CUD transfers to modern LLM regimes, we extend our evaluation to high-capacity decoder-only models using Llama-3.1-8B-Instruct as a teacher for a

Benchmark	Teacher (8B-inst)	Student (1B-inst)	LKD (Baseline)	Ours (Full)
MMLU	68.02	46.00	47.09	47.24
CSQA	77.31	54.95	55.20	58.56
HellaSwag	79.18	60.78	61.78	61.91
TruthfulQA	54.04	43.77	44.85	45.62
Avg	69.64	51.38	52.23	53.33

Table 7: Scalability results on knowledge distillation from Llama-3.1-8B-Instruct to Llama-3.2-1B-Instruct, evaluated on multiple-choice benchmarks via the lm-eval harness. CUD consistently outperforms the standard LKD baseline across all tasks.

Llama-3.2-1B-Instruct student. As shown in Table 7, CUD consistently outperforms the standard LKD baseline, achieving a higher average accuracy of **53.33%** (vs. 52.23% for LKD).

We mitigate the complexity of the vast $\sim 128K$ vocabulary by applying CUD exclusively to candidate answer tokens, thereby restricting the calibration to a task-relevant subset. This ensures our mechanism remains computationally efficient and strictly aligned with the log-likelihood-based scoring protocol. Such alignment provides initial evidence that CUD extends to decoder-only LLMs in the multiple-choice log-likelihood setting; further details are provided in Appendix F.

6 Related Work

Knowledge distillation (KD) fundamentally relies on transferring the “knowledge” encoded in soft targets (Hinton et al., 2015), yet standard teachers often exhibit overconfidence that degrades this signal. While calibration techniques (Guo et al., 2017) and regularizers like label smoothing (Müller et al., 2019) or confidence penalties (Pereyra et al., 2017) address predictive sharpness, they risk indiscriminately blurring the fine-grained inter-class geometry essential for effective transfer. Self- and born-again variants (Mobahi et al., 2020; Zhang et al., 2019) improve students without an external teacher, but do not explicitly preserve difficulty-aware uncertainty or prevent wrong-peak propagation. Furthermore, prior attempts to optimize teachers for distillation (Menon et al., 2021; Dong et al., 2024) or align intermediate features (Sun et al., 2019; Jiao et al., 2020) often lack difficulty-awareness or impose strict architectural constraints: feature alignment requires the student’s width and depth to be mappable to the teacher, which fails when both are reduced. Orthogonally, focal loss (Lin et al., 2017) down-weights trivially correct examples and

emphasizes hard ones.

In contrast, our approach targets the *distributional shape* of the transfer and remains architecture-agnostic; we jointly optimize the teacher to maintain difficulty-aware uncertainty through selective entropy shaping (DUS) and apply a constrained projection (W-Clip) to suppress specific wrong-class peaks without distorting the informative pairwise relations among background classes.

Recent work by Hamidi et al. (2024) also reformulates teacher training to better align predictions with the latent data distribution. While sharing the goal of optimizing teachers, our approach specifically targets informational calibration rather than distributional matching; we couple a focal term with a gated entropy reward to build a *distillation-friendly* teacher that increases uncertainty *only* where it is epistemic, unlike global smoothing or pure accuracy-driven shaping.

For an extended review of related literature, please refer to the Appendix L.

7 Conclusion

We presented CUD, a simple architecture-agnostic dual distillation mechanism that reshapes the teacher to retain difficulty-aware uncertainty and suppress over-confident mistakes while preserving informative class relations. Empirically, CUD outperforms strong baselines, with gains scaling with class cardinality and concentrated in high-cardinality settings. These results suggest that effective distillation depends not only on *how* knowledge is transferred, but also on *what* distributional structure the teacher provides.

Limitations

Our work has several limitations. First, we evaluate only single-label classification on standard benchmarks; multi-label tasks, distribution shift, and resource-limited scenarios remain open. Second, our projection rules use fixed hyperparameters (entropy weights, wrong-mass budget) that may not adapt optimally across domains or dynamic training conditions. Third, the benefits of calibrated uncertainty are reduced on binary or very small-class problems, where the teacher’s uncertainty structure becomes nearly one-dimensional and suppressing a wrong peak is almost equivalent to mild logit re-tempering. Lastly, we do not address open-ended generation. We observe gains on

decoder-only models under log-likelihood scoring over a fixed candidate set, but W-Clip assumes that a wrong top-1 class can be identified, whereas at each decoding step several continuations are acceptable and a token differing from the reference is not necessarily an error. These limitations suggest directions for future work: adaptive per-example calibration, integration with semantic neighborhoods or taxonomies, and extending DUS to open-ended generation.

Acknowledgements

This research was supported by a grant from the Korean ARPA-H Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (No. RS-2025-25456780); by grants from the National Research Foundation of Korea (NRF), funded by the Korean government (MSIT and MOE: RS-2025-16063688; MSIT: RS-2025-02215813, RS-2026-25491306); and by the Ewha Global Excellence Program Research Grant.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd workshop on natural language processing for conversational AI*, pages 38–45.
- Hyung Gun Chi, Florian Pesce, Wonil Chang, Oggi Rudovic, Arturo Argueta, Stefan Braun, Vineet Garg, and Ahmed Hussen Abdelaziz. 2025. Adaptive knowledge distillation for device-directed speech detection. *arXiv preprint arXiv:2508.02801*.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, and 1 others. 2023. Palm: Scaling language modeling with pathways. *Journal of machine learning research*, 24(240):1–113.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186.
- Chengyu Dong, Liyuan Liu, and Jingbo Shang. 2024. Toward student-oriented teacher network training for knowledge distillation. In *International Conference on Learning Representations*, volume 2024, pages 37191–37213.
- Maelaynayn El Baida, Farid Boushaba, and Mimoun Chourak. 2025. Drowning overconfidence with uncertainty: mitigating deep learning overconfidence in flood depth super-resolution through maximum entropy regularization. *Stochastic Environmental Research and Risk Assessment*, 39(3):1159–1177.
- Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, and 1 others. 2023. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302.
- Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, and 1 others. 2021. A framework for few-shot language model evaluation. *Zenodo*.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.
- Shayan Mohajer Hamidi, Xizhen Deng, Renhao Tan, Linfeng Ye, and Ahmed Hussein Salamah. 2024. How to train the teacher model for effective knowledge distillation. In *European Conference on Computer Vision*, pages 1–18. Springer.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. 2022. Knowledge distillation from a stronger teacher. *Advances in neural information processing systems*, 35:33716–33727.
- Wenqi Jiang, Shuai Zhang, Boran Han, Jie Wang, Bernie Wang, and Tim Kraska. 2025. Piperag: Fast retrieval-augmented generation via adaptive pipeline parallelism. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 589–600.
- Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling bert for natural language understanding. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 4163–4174.

- Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, and 1 others. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1311–1316.
- Xin Li and Dan Roth. 2002. Learning question classifiers. In *Coling 2002: The 19th international conference on computational linguistics*.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 3214–3252.
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.
- Chang Liu, Chongyang Tao, Jiazhan Feng, and Dongyan Zhao. 2022. Multi-granularity structural knowledge distillation for language model compression. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1001–1011.
- Aditya K Menon, Ankit Singh Rawat, Sashank Reddi, Seungyeon Kim, and Sanjiv Kumar. 2021. A statistical perspective on distillation. In *International Conference on Machine Learning*, pages 7632–7642. PMLR.
- Ishan Mishra, Riyanshu Jain, Dhruv Viradiya, Divyam Patel, and Deepak Mishra. 2023. Knowledge distillation with ensemble calibration. In *Proceedings of the Fourteenth Indian Conference on Computer Vision, Graphics and Image Processing*, pages 1–10.
- Hossein Mobahi, Mehrdad Farajtabar, and Peter Bartlett. 2020. Self-distillation amplifies regularization in hilbert space. *Advances in Neural Information Processing Systems*, 33:3351–3361.
- Amir Moslemi, Anna Briskina, Zubeka Dang, and Jason Li. 2024. A survey on knowledge distillation: Recent advancements. *Machine Learning with Applications*, 18:100605.
- Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? *Advances in neural information processing systems*, 32.
- Geondo Park, Gyeongman Kim, and Eunho Yang. 2021. Distilling linguistic context for language model compression. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 364–378.
- Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*.
- Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient knowledge distillation for bert model compression. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 4323–4332.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. 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)*, pages 4149–4158.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. 2022. Mitigating neural network overconfidence with logit normalization. In *International conference on machine learning*, pages 23631–23644. PMLR.
- Siyue Wu, Hongzhan Chen, Xiaojun Quan, Qifan Wang, and Rui Wang. 2023. Ad-kd: Attribution-driven knowledge distillation for language model compression. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 8449–8465.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4791–4800.
- Lin Feng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. 2019. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3712–3721. IEEE.
- Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

A Theoretical Analysis of the Constrained Projection

In this appendix, we provide the rigorous mathematical justification for the projection problem defined in Section 2, addressing the existence and uniqueness of the solution and deriving the connection between the optimal solution and our proposed modules.

A.1 Existence and Uniqueness of the Solution

R1 comprises two conditions of different character: an entropy bound, which is not linear in q , and a lower bound on the mass over $\mathcal{N}(y; x)$, which is. CUD handles them in two stages. The entropy bound is imposed during teacher training as a soft constraint through the focal-entropy objective of Eq. (3), so no projection onto an entropy level set is ever computed; the projection of Eq. (2) is then applied to the remaining linear constraints, namely the neighbourhood lower bound and R2. The analysis below concerns that projection.

Theorem 1 (Uniqueness of Projection). *Let $\mathcal{Q}$ be the set of valid probability distributions on the simplex Δ^K . The feasible set $\mathcal{C} = \{q \in \mathcal{Q} \mid \sum_{k \in \mathcal{N}} q_k \geq \alpha, \sum_{k \in \mathcal{W}} q_k \leq \varepsilon\}$ is defined by the intersection of linear half-spaces and the probability simplex, making $\mathcal{C}$ a convex polytope. Since the objective function $f(q) = D_{KL}(q \parallel p_T)$ is strictly convex with respect to q (for $q, p_T > 0$), and the feasible set $\mathcal{C}$ is convex and compact, standard convex optimization theory guarantees that if $\mathcal{C}$ is non-empty, there exists a **unique** global minimizer q^* .*

The above guarantees that our target calibration objective does not suffer from instability due to multiple local minima.

A.2 KKT Analysis and Exponential Tilting

To characterize the solution, we formulate the Lagrangian $\mathcal{L}$. Both linear constraints can be written generally as $\sum_{k \in S} q_k \leq C$. Introducing Lagrange multipliers μ for the neighbourhood lower bound and ν for the wrong-mass upper-bound constraint (R2), the KKT stationarity condition yields:

$$\frac{\partial \mathcal{L}}{\partial q_k} = \log q_k - \log p_T(k) + 1 + \sum_i \lambda_i \frac{\partial \text{Constraints}}{\partial q_k} = 0 \quad (11)$$

Solving for q_k , we obtain the closed-form solution:

$$q_k^* = \frac{1}{Z} p_T(k) \exp(\mu \cdot \mathbb{I}(k \in \mathcal{N}) - \nu \cdot \mathbb{I}(k \in \mathcal{W})) \quad (12)$$

where Z is the normalization constant. This confirms that the optimal projection is an *exponential tilting* of the teacher distribution, scaling the probabilities in the constrained sets $\mathcal{N}$ and $\mathcal{W}$ while preserving the relative ratios of probabilities for unconstrained classes (the “dark knowledge”).

A.3 W-Clip as a First-Order Approximation

We now show that the W-Clip operation approximates the optimal solution derived in Eq. (12). Consider the case where the wrong-mass constraint is active for the top-1 class k^* , i.e., $\mathcal{W} = \{k^*\}$. The optimal update requires scaling $p_T(k^*)$ by a factor $e^{-\nu}$.

The budget in Eq. (5) caps the tilt at $\nu = \delta(x)/p_T(k^* \mid x) \leq \eta$, so the expansion is only ever evaluated on $\nu \in (0, \eta]$. Performing a first-order Taylor expansion around $\nu = 0$:

$$e^{-\nu} \approx 1 - \nu \quad (13)$$

Substituting this into the update rule:

$$q_{k^*}^* \approx \frac{1}{Z} p_T(k^*) (1 - \nu) \approx p_T(k^*) - \nu p_T(k^*) \quad (14)$$

This linear subtraction form is functionally equivalent to the W-Clip operation defined in Eq. (8), where the clipped mass $\delta(x)$ corresponds to the term $\nu p_T(k^*)$. Thus, W-Clip provides a computationally efficient, first-order approximation to the theoretically optimal exponential projection, avoiding the need for iterative solvers during the training loop.

Over this range the linear factor $(1 - \nu)$ departs from $e^{-\nu}$ by at most 17.6% in relative terms, and by only 0.2% in the small-margin regime where the wrong top-1 is a semantic neighbour of the ground truth (cf. Table 1).

Proposition 1 (Budgeted suppression). *Let $k^* \neq y$ and let $\tilde{p}_T$ be given by Eqs. (5)–(8) with $\eta \in (0, 1)$. Then*

$$(1 - \eta) p_T(k^* \mid x) \leq \tilde{p}_T(k^* \mid x) \leq p_T(k^* \mid x).$$

The upper bound says that mass on the wrong top-1 always decreases, in the direction R2 intends, with stronger suppression the more confidently the teacher places the wrong class above the ground

Hyperparameter	Value	Accuracy
λ_H	0.1	91.22
	0.2	91.16
	0.3	91.04
	0.4	91.13
m	0.1	91.21
	0.3	91.22
	0.5	91.22
	0.7	91.23
	0.9	91.22
η	0.1	91.19
	0.3	91.22
	0.5	91.23
	0.7	91.22
	0.9	91.24
$\lambda_{KD}/\lambda_{CE}$	0.2 / 0.8	90.73
	0.4 / 0.6	91.04
	0.6 / 0.4	91.21
	0.8 / 0.2	91.22

Table 8: **Hyperparameter sensitivity results.** Bold values indicate the configuration used in the main experiments.

truth. The lower bound says that the removal is capped at η , so the wrong peak is never erased entirely and at least a $(1 - \eta)$ fraction is retained.

B Hyperparameter Sensitivity Analysis

To better understand the robustness of our hyperparameter choices, we conduct a sensitivity study across four key parameters. Bold values in each table correspond to the settings used for our main experimental results. As shown, CUD exhibits low sensitivity across a broad range of configurations, further justifying our decision to perform minimal per-dataset tuning.

C Detailed Benchmark Description

We evaluate our method on the following benchmark datasets:

Banking77 (77 classes, Customer Support, Intent Classification). A fine-grained intent dataset for banking-related customer queries (Casanueva et al., 2020).

CLINC150 (150 classes, Open-domain, Intent Classification). A wide-coverage natural language

intent dataset spanning 150 intent classes across multiple domains (Larson et al., 2019).

MASSIVE (60 classes, Multilingual SLU, Intent Classification). A multilingual dataset for spoken language understanding; we use the English subset unless otherwise stated (FitzGerald et al., 2023).

TREC (6 classes, Question Answering, Question Type Classification). A benchmark dataset with coarse-grained question type labels for factoid QA tasks (Li and Roth, 2002).

AG News (4 classes, News, Topic Classification). A news article classification dataset with four broad topical categories (Zhang et al., 2015).

We provide here a metadata summary of all benchmark datasets used in our experiments. This section outlines class cardinalities, domains, and task types, offering additional context for the diversity of evaluation settings.

D On Binary Classification

Method	QNLI	SST-2
FT Teacher	90.66	90.82
LKD	89.09	90.71
PKD	89.65	90.82
TinyBERT [†]	90.50	91.60
AD-KD	89.84	91.62
CUD (Ours)	89.23	91.28

Table 9: **Binary-task comparison on QNLI and SST-2** with a 6-layer, 768-dim student. [†] Methods are trained with distillation already at the pretraining stage; we fine-tune publicly available checkpoints.

As discussed in the main text, calibrated uncertainty distillation is most effective when the teacher distribution contains rich multi-class structure, which provides informative relative probabilities over many plausible alternatives. In contrast, for binary or near-binary tasks, the teacher’s output reduces to a one-dimensional distribution of the form $(p, 1 - p)$, limiting the amount of dark knowledge available for transfer.

To verify this empirically, we report results on two standard binary GLUE tasks, QNLI and SST-2. Table 9 compares CUD against representative KD baselines on two binary GLUE tasks. Consistent with our analysis, the gains from calibrated uncertainty are substantially smaller than on high-cardinality tasks (e.g., Banking77 or CLINC150).

In some cases, the performance is nearly identical to LKD, reflecting that in the binary case there are no other classes across which to preserve relative structure, so suppressing a wrong peak provides little information beyond the hard label.

Overall, these findings reinforce a key property of calibrated uncertainty distillation: its advantages grow with class cardinality and class ambiguity. In binary tasks—where the teacher’s uncertainty structure collapses to a single degree of freedom—the potential benefit of redistributing probability mass is inherently limited. Nevertheless, our method remains competitive and does not degrade accuracy, indicating that CUD is a safe replacement for conventional KD even in low-cardinality settings.

E Additional Results on Uncertainty Quantification

Task	Method	ECE	Brier
Banking77	TinyBERT	0.076	0.203
	CKD	0.047	0.159
	AD-KD	0.039	0.112
	CUD (Ours)	0.359	0.237
CLINC150	TinyBERT	0.066	0.203
	CKD	0.028	0.067
	AD-KD	0.021	0.068
	CUD (Ours)	0.394	0.228
TREC	TinyBERT	0.013	0.061
	CKD	0.023	0.048
	AD-KD	0.021	0.047
	CUD (Ours)	0.182	0.099
AGNews	TinyBERT	0.009	0.095
	CKD	0.047	0.099
	AD-KD	0.027	0.086
	CUD (Ours)	0.061	0.095
MASSIVE	TinyBERT	0.059	0.188
	CKD	0.080	0.223
	AD-KD	0.074	0.182
	CUD (Ours)	0.101	0.176

Table 10: ECE and Brier calibration performance across tasks. Lower values indicate better calibration.

In this section, we provide a more granular analysis of the uncertainty estimates, specifically focusing on the behavioral shifts and practical implications for selective prediction.

Dataset	Teacher	Acc	ECE	Brier	ECE_{wrong}	$Brier_{\text{wrong}}$
Banking77	FT	93.70	0.045	0.120	0.817	1.654
	DUS	92.43	0.172	0.170	0.546	1.173
CLINC150	FT	96.35	0.025	0.065	0.790	1.601
	DUS	94.93	0.049	0.065	0.619	1.275
TREC	FT	97.40	0.035	0.067	0.933	1.790
	DUS	97.00	0.031	0.070	0.924	1.751
AGNews	FT	94.69	0.024	0.114	0.801	1.406
	DUS	94.72	0.219	0.218	0.504	0.743
MASSIVE	FT	89.47	0.075	0.186	0.813	1.641
	DUS	89.47	0.070	0.193	0.623	1.269

Table 11: **Full teacher calibration results.** FT is the standard fine-tuned teacher, DUS the teacher trained with our objective. Overall ECE increases on some datasets, most visibly on AGNews: DUS lowers confidence across the board, producing the controlled underconfident shift discussed above. Calibration error restricted to the teacher’s own mistakes improves on every dataset.

E.1 Calibration Quality and Behavioral Shifts

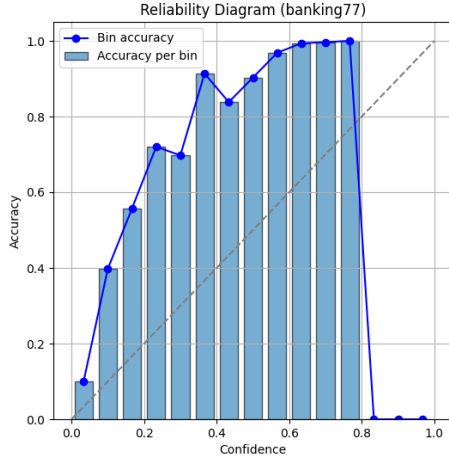
To evaluate how well each model signals its own mistakes, we report ECE_{wrong} and $Brier_{\text{wrong}}$ in Table 12, computed strictly over incorrect predictions. CUD achieves markedly lower values on four of five datasets. This indicates that the student inherits the ability to assign lower confidence when it is incorrect, successfully mitigating the inherent overconfidence issue; we verify the corresponding behavior on the teacher itself below.

Furthermore, across all datasets, DUS reduces overconfident predictions by reshaping the teacher distribution, leading to a controlled underconfident shift. Although this can technically increase overall ECE in cases where confidence is lower than accuracy, we emphasize that this behavior is inherently **safer than overconfidence**. This shift yields more reliable uncertainty estimates for downstream OOD detection and risk-sensitive settings, where failing to signal a mistake is more costly than being cautious.

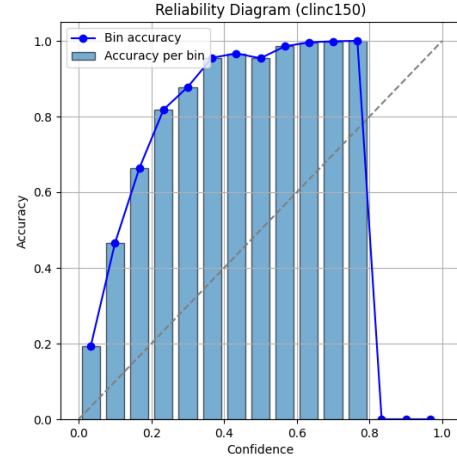
Table 11 reports these metrics for the teacher itself. The same pattern holds one level up: overall ECE rises while calibration error over the teacher’s own mistakes falls on every dataset, with the smallest change on TREC, whose teacher errs on only 2.6% of test examples.

E.2 Correctness Separation and Selective Prediction

Beyond point-wise calibration, we assess whether the generated confidence scores can effectively separate correct from incorrect predictions.



(a) Reliability diagram on Banking77.



(b) Reliability diagram on Clinc150.

Figure 3: **Effect of DUS on predictive calibration.** Across datasets, DUS reduces overconfident predictions by reshaping the teacher distribution, leading to a controlled underconfident shift. Although this increases ECE due to lower confidence than accuracy, the behavior is safer than overconfidence and yields more reliable uncertainty estimates for downstream OOD detection and risk-sensitive settings.

ROC Analysis. As shown in Figure 4a, the ROC curves (where the label equals prediction correctness) on BANKING77 reveal a visibly steeper rise in the low False Positive Rate (FPR) region. This demonstrates that CUD-calibrated targets produce confidence scores that better rank easy and hard examples. Such ranking performance is critical for **selective prediction**, where a model must decide whether to commit to a prediction or defer to a human expert.

Ablation on Separation Performance. Individual module analysis confirms that both DUS and W-Clip provide complementary inductive biases for uncertainty estimation. Individually, DUS and W-Clip improve the correctness AUROC over the vanilla LKD baseline (attaining 0.90 and 0.88, respectively, compared to 0.83). Their combination in CUD yields the strongest separation of 0.92, ensuring the most robust uncertainty estimates for practical deployment.

F Llama Distillation Details

F.1 Datasets

Training Data. For teacher fine-tuning, we use the MMLU auxiliary_train split (Hendrycks et al., 2020) (99,842 samples), formatted in the lm-eval multiple-choice style.

For student distillation, we convert the same split into an instruction-following SFT format, where

Task	Method	ECE _{wrong}	Brier _{wrong}
Banking77	TinyBERT	0.858	1.696
	CKD	0.747	1.481
	AD-KD	0.791	1.557
	CUD (Ours)	0.277	1.033
CLINC150	TinyBERT	0.861	1.741
	CKD	0.835	1.657
	AD-KD	0.730	1.480
	CUD (Ours)	0.236	1.023
TREC	TinyBERT	0.929	1.851
	CKD	0.870	1.675
	AD-KD	0.950	1.900
	CUD (Ours)	0.566	0.993
AGNews	TinyBERT	0.794	1.393
	CKD	0.972	1.919
	AD-KD	0.865	1.570
	CUD (Ours)	0.801	1.396
MASSIVE	TinyBERT	0.692	1.469
	CKD	0.764	1.540
	AD-KD	0.833	1.667
	CUD (Ours)	0.586	1.213

Table 12: ECE_{wrong} and Brier_{wrong} calibration performance across tasks on wrong predictions.

each sample consists of a question with labeled choices and the correct answer letter as the target output. Using the same questions in both stages ensures that every distillation sample carries a ground-truth answer, which W-Clip requires to identify a wrong top-1 prediction.

Evaluation Data. We evaluate on four benchmarks using the EleutherAI 1m-eval harness (Gao et al., 2021):

- **MMLU** (Hendrycks et al., 2020): 57-subject 4-choice QA (acc, test split)
- **CommonsenseQA** (Talmor et al., 2019): 5-choice commonsense reasoning (acc, validation split)
- **HellaSwag** (Zellers et al., 2019): 4-choice sentence completion (acc_norm, validation split)
- **TruthfulQA** (Lin et al., 2022): multi-true multiple-choice (mc2)

F.2 Evaluation Protocol

Following the 1m-eval protocol, we measure the conditional log-likelihood of each answer choice given the context:

$$\log P(\text{choice}_i \mid \text{context}) = \sum_t \log P(\text{token}_t \mid \text{context}, \text{token}_{<t}) \quad (15)$$

Rather than relying on open-ended generation over the full 128K-token vocabulary, this approach directly computes calibrated probabilities over a small set of candidate answers. The predicted answer is the choice with the highest log-likelihood (length-normalized for HellaSwag).

F.3 Adapting CUD to Decoder-Only LLMs

A key challenge in applying CUD to generative LLMs is the vocabulary size ($|V|=128\text{K}$): calibrating the full teacher distribution is computationally expensive and introduces noise from irrelevant tokens. We address this by adopting the log-likelihood-based multiple-choice scoring used in 1m-eval (Gao et al., 2021), which computes the conditional log-likelihood of each answer choice given the context and normalizes over the candidate set. This reduces the effective output space from $\sim 128\text{K}$ tokens to a handful of answer candidates (typically 4–5). Consequently, CUD’s calibration mechanisms—including W-Clipping—can operate directly on a compact, task-relevant probability distribution without modifying the remaining vocabulary.

Dataset	TS	LS	CUD (Ours)
Banking77	80.61	81.29	91.22
CLINC150	88.35	86.88	93.45
TREC	95.80	95.80	96.47
AGNews	94.31	94.19	94.11
Massive	82.04	82.19	87.22
Avg	88.22	88.07	92.49

Table 13: **Comparison of smoothing strategies.** Temperature Scaling (TS) and Label Smoothing (LS) rely on simple, rule-based smoothing, whereas our DUS-based method reshapes the teacher’s scores in a way that is more suitable for knowledge distillation.

G Comparison with TS and LS

We also compare our calibrated distillation against simpler smoothing baselines such as Temperature Scaling (TS) and Label Smoothing (LS). This comparison highlights that merely smoothing the teacher’s distribution—without modeling sample-dependent structure—does not preserve the rich information needed for effective distillation. Table 13 reports a direct comparison between TS, LS, and our approach under the same experimental setup.

All three methods are evaluated under an identical teacher–student architecture. The student is a 4-layer, 256-dimensional BERT model (google/bert_uncased_L-4_H-256_A-4). For each method, we perform a grid search over learning rates $\{5 \times 10^{-4}, 10^{-4}, 5 \times 10^{-5}\}$. All other training settings (optimizer, batch size, number of epochs, etc.) are kept fixed so that the only difference lies in how the teacher’s distribution is smoothed.

As shown in Table 13, our method consistently improves over TS and LS in most cases. The gains are especially notable on high-cardinality tasks such as Banking77, CLINC150, and MASSIVE. This suggests that, compared to simple rule-based methods that apply a uniform temperature or fixed label smoothing, our loss preserves richer dark knowledge by adapting the teacher distribution to each sample’s difficulty and uncertainty. In other words, our approach does more than merely “soften” the teacher: it reshapes the teacher’s confidence structure in a way that is more conducive to effective knowledge distillation.

Baseline	Dataset	FT Teacher	CUD (Ours)
TinyBERT	Banking77	89.74	91.29
	CLINC150	88.44	92.91
	TREC	96.20	96.80
	AGNews	94.31	94.36
	Massive	89.12	89.27
MGSKD	Banking77	91.62	91.92
	CLINC150	95.46	95.91
	TREC	87.80	97.20
	AGNews	94.81	94.44
	Massive	88.88	89.27
CKD	Banking77	93.76	93.86
	CLINC150	96.11	97.03
	TREC	97.80	97.80
	AGNews	94.89	94.73
	Massive	89.57	89.86
AD-KD	Banking77	93.05	93.08
	CLINC150	95.53	95.77
	TREC	96.40	97.60
	AGNews	94.82	94.55
	Massive	89.62	89.32

Table 14: Comparison of layer-6 student performance using a fine-tuned teacher vs. our CUD-trained teacher. Single-run results; the FT-teacher column therefore differs slightly from the three-run means in Table 2.

Dataset	FT Teacher	Ours ($\gamma=5$)	Ours ($\gamma=10$)
Banking77	93.70	91.62	92.43
CLINC150	96.35	94.17	94.93
TREC	97.40	97.40	97.00
AGNews	94.69	94.42	94.72
MASSIVE	89.47	89.12	89.47

Table 15: **Teacher performance under different levels of DUS smoothing.** The *FT Teacher* row reports the accuracy of a standard fine-tuned teacher, while *Ours* corresponds to the same architecture trained with DUS at different smoothing strengths γ .

H Additional Teacher side Ablations

For KD baselines, we also evaluate the effect of replacing the conventional fine-tuned teacher with our CUD-trained teacher while keeping the student architecture and loss fixed. As shown in Table 14, almost all baselines and benchmarks exhibit consistent performance gains, indicating that substituting the standard teacher with our calibrated teacher provides clear benefits.

This suggests that, although the proposed teacher training procedure does not necessarily improve (and can even slightly reduce) the teacher’s raw accuracy (cf. Table 15), it yields a teacher that encodes richer dark knowledge about inter-class relationships. In other words, even under the same distillation architecture, the quality of information

Method	Wall-clock Time	Relative Cost
Standard FT Teacher	11m 06s (665.8 s)	1.00×
CUD Teacher (Ours)	12m 10s (729.8 s)	1.10×

Table 16: **Computation cost comparison during teacher training.** CUD incurs only a small overhead of approximately 10% relative to standard fine-tuning.

transferred to the student depends strongly on how the teacher itself is trained, and our teacher provides more informative soft targets.

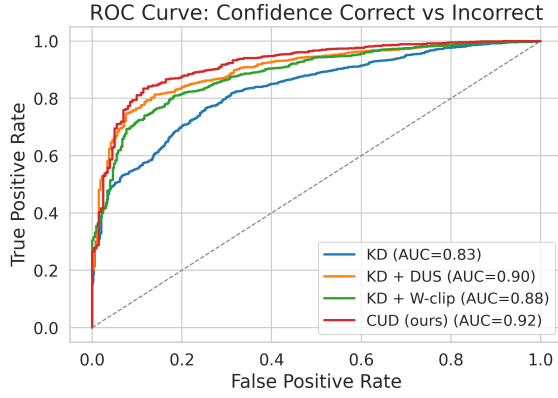
The effect is particularly pronounced on datasets with a large number of classes, such as Banking77, CLINC150, and MASSIVE. In these multi-class settings, decision boundaries and error patterns become more complex, so the teacher’s ability to capture subtle relationships among classes has a direct impact on distillation performance. Our teacher better models the underlying uncertainty and correlations over this complex label space, thereby increasing the amount of useful information available to the student through soft targets.

By contrast, on datasets with relatively few classes, such as TREC and AGNews, the amount of exploitable dark knowledge is inherently limited. Consequently, the absolute gains from emphasizing this structure are smaller. Nevertheless, our teacher still matches or slightly surpasses the standard fine-tuned teacher on most of these tasks. Taken together, these results indicate that the proposed teacher training strategy yields substantial benefits in challenging multi-class regimes, while remaining a safe and stable choice in simpler, low-cardinality settings.

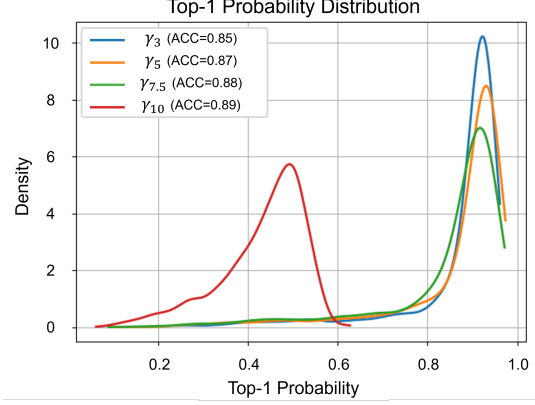
I Computation Analysis

DUS introduces minimal computational overhead because it modifies only the loss function while leaving the model architecture and training pipeline unchanged. Theoretical FLOPs per iteration are therefore identical to those of standard fine-tuning.

It is important to note that our method—like the standard practice in response-based KD (e.g., PKD, TinyBERT, CKD, AD-KD, MGSKD)—assumes a task-specific fine-tuned teacher. Although the cost of obtaining such a teacher may appear “hidden” in real-world scenarios, where many pre-tuned large models are publicly available, this requirement is shared across nearly all KD pipelines. Under the assumption that a task-specific fine-tuned teacher—one that is not publicly available—must



(a) ROC curves comparing confidence-based correctness separation.



(b) Teacher probability distributions under varying γ . Higher γ values soften over-confident peaks.

Figure 4: Effect of calibration on uncertainty and distributions.

be trained for each task, the only additional computation introduced by DUS arises from the modified loss.

Table 16 reports the wall-clock time on BANKING77: DUS teacher training takes 12m 10s, compared to 11m 6s for standard fine-tuning, corresponding to an overhead of only 10%. The student-side distillation cost remains essentially unchanged, as the complexity of our loss is comparable to that of standard KL-based KD objectives.

Overall, DUS provides a meaningful perspective on knowledge distillation by reshaping the teacher into a more calibrated and informative distribution, while adding only negligible computational cost, even when accounting for the teacher fine-tuning step.

J The Role of γ : Balancing accuracy and Calibrated uncertainty

We study how γ in DUS reshapes the teacher distribution and student performance. Since the per-example weight scales as $(1 - p_T^y)^\gamma$, increasing γ aggressively down-weights correct examples and emphasizes wrong ones. Empirically (Fig. 4b), larger γ produces a *softer* top-1 distribution: the sharp peak near ≈ 0.95 collapses while probability mass shifts toward mid confidence (0.5–0.9). Thus, there is a clear **inverse relation** between γ and top-1 concentration/peak height (and, equivalently, a positive relation with entropy on hard cases). This softening is beneficial for transfer up to a point. On the dev sweep shown, student accuracy rises monotonically as γ increases from 3 to 10 (0.85→0.89; legend of Fig. 4b), indicating that

within this range recovering teacher uncertainty improves the ranking signal distilled to the student. Although $\gamma=10$ exceeds the $\gamma \approx 2$ standard in focal loss (Lin et al., 2017), our focal factor *gates the entropy reward* rather than reweighting classification loss, so a more aggressive γ is needed to fully suppress the reward on confident-correct samples. In our experiments, the γ sweep further illustrates the transfer-first trade-off: modest increases in teacher uncertainty (sometimes with a minor drop in the teacher’s standalone accuracy) correlate with *higher* student accuracy after KD, reinforcing that calibrating the teacher for *distillability* can be preferable to maximizing its peak accuracy.

K Learning Curve

Figure 5 reports the learning curves comparing our method. The training loss of our method decreases smoothly and stabilizes early, demonstrating that the optimization process is well-behaved even though its absolute loss scale differs from AD-KD due to the different loss terms used.

Validation accuracy improves steadily and remains consistently higher than that of AD-KD throughout training, reaching a stable plateau without late-stage degradation.

Overall, these learning curves confirm that (1) our optimization process converges reliably, and (2) the performance gains are attributable to a better-shaped training objective rather than incidental optimization effects.

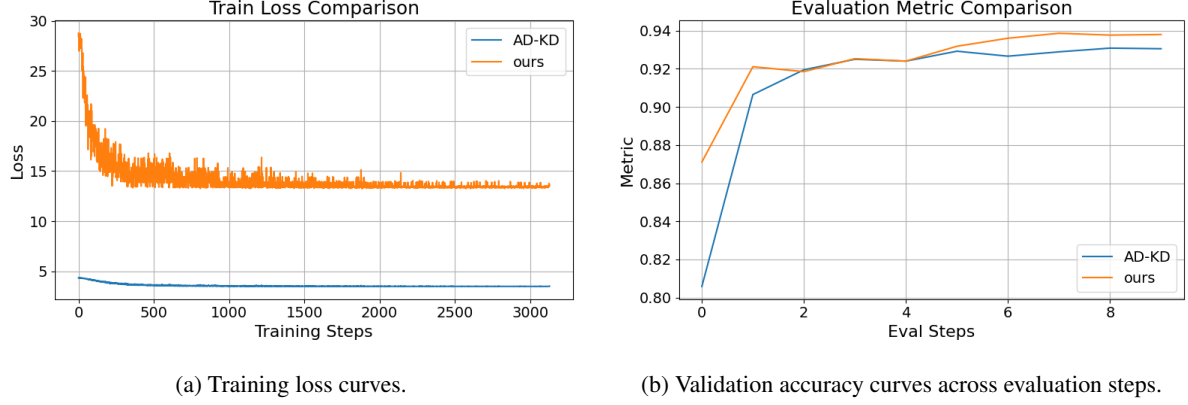


Figure 5: Learning curves illustrating the convergence and generalization behavior of our method. Our approach demonstrates smoother optimization and stronger validation performance throughout training.

L Extended Related Work

KD under uncertainty, calibration, and class structure. Classical response-based KD transfers a teacher’s softened output distribution to the student (Hinton et al., 2015), leveraging “dark knowledge”—relative probabilities over non-ground-truth classes—which is especially informative when class cardinality is large. A parallel literature studies calibration and uncertainty of modern networks (Guo et al., 2017), and regularizers that alter predictive sharpness, such as label smoothing (Müller et al., 2019) and confidence penalties (Pereyra et al., 2017). While these techniques can improve expected calibration, they may also blur meaningful inter-class geometry that KD intends to pass on. Self-/born-again distillation variants (Mobahi et al., 2020; Zhang et al., 2019) improve students without an external teacher, but they do not explicitly preserve difficulty-aware uncertainty or prevent wrong-peak propagation. Our approach targets the *distributional shape* transferred by KD: we (i) raise teacher entropy selectively on hard inputs to keep informative neighborhoods (R1), and (ii) project the teacher outputs to respect a wrong-mass budget (R2), thus preserving pairwise odds for the untouched classes.

Teacher shaping for distillation. Beyond treating the teacher as fixed, several works revisit how teachers should be trained for better students. (Menon et al., 2021) argue that soft targets closer to the latent label distribution yield stronger generalization than one-hot supervision; (Dong et al., 2024) further reformulate teacher training with regularizers that move predictions toward a faithful class distribution for the data. Orthogonally, focal

loss (Lin et al., 2017) down-weights trivially correct examples and emphasizes hard ones; we adapt this idea to build a *distillation-friendly* teacher by coupling a focal term with a gated entropy reward, thereby increasing uncertainty *only* where the teacher is wrong or uncertain, unlike global smoothing.

Selective imitation and error-aware KD. Confidence-aware distillation methods reweight or filter teacher supervision based on teacher confidence, aiming to reduce the harm of overconfident mistakes while keeping the simplicity of response-based KD (e.g., “adaptive” or “selective” KD variants). Feature-based approaches such as PKD (Sun et al., 2019) and TinyBERT (Jiao et al., 2020) align intermediate representations (hidden states, attention) and can be very effective when student width/depth can be mapped to the teacher; however, they are less applicable under severe architectural mismatch (e.g., depth *and* width changes). Our method is architecture-agnostic: rather than aligning features, we enforce a small *projection* on the teacher distribution that (a) suppresses only the wrong top-1 mass under a budget (R2) and (b) leaves the rest of the class relations intact, which theory and our results indicate is most beneficial as the number of classes grows.

Relative to prior KD, our contribution is to *jointly* (1) shape the teacher to preserve difficulty-aware uncertainty and (2) calibrate the transfer via a constrained projection that prevents error propagation. This coupling operationalizes the intuition that KD helps most when the teacher behaves like a calibrated posterior on both easy and hard inputs, and explains the empirical trend that our gains in-

crease with class cardinality while becoming comparable to temperature scaling or smoothing on binary tasks.

M Algorithm

Algorithm 1 Proposed Knowledge Distillation Procedure

Require: Teacher T , Student S , W-Clip budget η , margin scale m

Step 1: Teacher Training (DUS)

for epoch **do**

$T(x) \leftarrow$ teacher logits

$\mathcal{L}_{CE} \leftarrow \text{CrossEntropy}(T(x), y)$

$\mathcal{L}_{focal} \leftarrow$ focal loss on $(T(x), y)$

$\mathcal{L}_{ent} \leftarrow$ entropy loss on wrong samples

$\mathcal{L}_{total} \leftarrow \mathcal{L}_{CE} + \mathcal{L}_{focal} + \mathcal{L}_{ent}$

Update T with $\mathcal{L}_{total}$

end for

Step 2: Student Distillation with W-Clip

for epoch **do**

$S(x) \leftarrow$ student logits

$p_T(\cdot | x) \leftarrow \text{softmax}(T(x))$

$k^* \leftarrow \arg \max_k p_T(k | x)$

if $k^* = y$ **then** $\triangleright$ teacher correct: bypass

W-Clip

$\tilde{p}_T(\cdot | x) \leftarrow p_T(\cdot | x)$

else $\triangleright$ teacher wrong: apply W-Clip

$\delta_{\text{budget}} \leftarrow \eta \cdot p_T(k^* | x)$

$\delta_{\text{margin}} \leftarrow m \cdot (p_T(k^* | x) - p_T(y | x))$

$\delta \leftarrow \min\{\delta_{\text{budget}}, \delta_{\text{margin}}\}$

$\tilde{p}_T(y | x) \leftarrow p_T(y | x) + \delta$

$\tilde{p}_T(k^* | x) \leftarrow p_T(k^* | x) - \delta$

$\tilde{p}_T(k | x) \leftarrow p_T(k | x)$ for $k \notin \{y, k^*\}$

end if

$\mathcal{L}_{CE} \leftarrow \text{CrossEntropy}(S(x), y)$

$\mathcal{L}_{KD} \leftarrow \text{KL}(S(x) \parallel \tilde{p}_T(\cdot | x))$

$\mathcal{L}_{total} \leftarrow \mathcal{L}_{CE} + \mathcal{L}_{KD}$

Update S with $\mathcal{L}_{total}$

end for
