

Calibrating Student Models via Uncertainty-Integrated Distillation

Ruixuan Zhang¹, Hongye Tan^{1,*}, Zhiwei Hu², Xiaoqi Han³,
Ru Li¹, Jeff Z. Pan⁴, Víctor Gutiérrez-Basulto^{5,*}

¹School of Computer and Information Technology, Shanxi University, China

²College of Information Science and Engineering, Shanxi Agricultural University, China

³College of Artificial Intelligence, Taiyuan University of Technology, China

⁴School of Informatics, University of Edinburgh, UK

⁵School of Computer Science and Informatics, Cardiff University, UK

*Correspondence: tanhongye@sxu.edu.cn, gutierrezbasulto@cardiff.ac.uk

Abstract

Model calibration aligns a model’s confidence with the accuracy of its predictions, directly influencing reliability and user trust. Surprisingly, existing research on knowledge distillation has paid little attention to the calibration of student models. In this paper, we evaluate calibration across diverse knowledge distillation settings and demonstrate that student models distilled from black-box teachers generally exhibit poor calibration. We attribute this to the fact that distillation fails to convey well-calibrated confidence, which serves as a crucial expression of predictive uncertainty. To address this limitation, we propose an *Uncertainty-Integrated Knowledge Distillation (UIKD)* method comprising two key components: (1) *uncertainty estimation*, which leverages the collective intelligence of multiple LLMs via a debate mechanism to quantify uncertainty in the distillation data; and (2) *uncertainty integration*, which employs an adaptive weighted loss function to integrate this data uncertainty into the student model effectively. Experiments on common-sense reasoning and natural language inference benchmarks demonstrate that UIKD consistently improves student model calibration while maintaining predictive accuracy. Our code is available at <https://github.com/RuixuanZhang127/UIKD>.

1 Introduction

While large language models (LLMs) demonstrate exceptional performance across diverse downstream tasks (Team, 2024; Yang et al., 2025a; DeepSeek-AI et al., 2025; Xu et al., 2025; Yan et al., 2025), their prohibitive computational and energy costs limit widespread deployment. Knowledge distillation (KD) addresses this bottleneck by transferring capabilities from massive teacher models to compact student models (Yang et al., 2025b), significantly reducing resource overhead. Crucially, for student models to be practically useful and safe,

we must look beyond accuracy to calibration: *the alignment between a model’s confidence and its actual correctness*. As illustrated in Figure 1, a well-calibrated model serves as a trustworthy indicator of reliability. Conversely, poor calibration poses severe risks in high-stakes environments; for instance, a medical LLM that assigns high confidence to an erroneous diagnosis could lead to treatment errors.

Although the literature abounds with methods to enhance student accuracy (Hsieh et al., 2023; Jiang et al., 2025; Chen et al., 2025; Xia et al., 2025a), the calibration of these distilled models remains significantly underexplored. Shum et al. (2024) attempt to mitigate miscalibration via temperature scaling on teacher logits. Yet, this approach necessitates access to the teacher’s internal probability distributions, rendering it unsuitable for black-box models where such confidence scores are inaccessible, a critical restriction given the ubiquity of proprietary LLMs. Moreover, existing literature lacks a systematic analysis of how diverse knowledge distillation settings impact student calibration.

In this paper, we systematically analyze the calibration of student models through both empirical evaluation and theoretical insights. Our analysis spans diverse KD settings, including white- and black-box paradigms, as well as varying model scales and architectures. We observe that student models distilled from black-box teachers consistently exhibit inferior calibration compared to those distilled from white-box teachers. Further analysis reveals that reliable uncertainty signals, derived from well-calibrated teacher confidence, are essential for inducing student calibration. However, black-box distillation inherently obscures these confidence scores, thereby impeding the effective transmission of uncertainty information.

To address this challenge, we propose *Uncertainty-Integrated Knowledge Distillation (UIKD)*, a framework specifically designed for

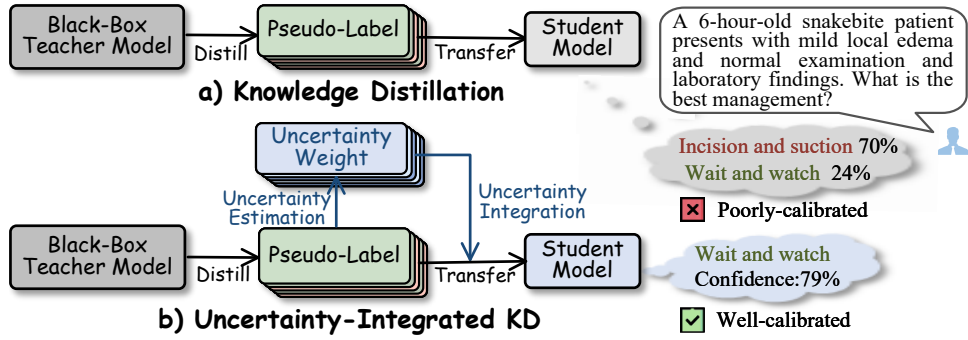


Figure 1: Comparison between (a) traditional KD and (b) Uncertainty-Integrated KD.

black-box models, cf. Figure 1(b). To overcome black-box constraints that preclude access to internal confidence scores, we estimate data uncertainty through multi-round debates among independent LLMs. This approach triggers iterative discourse only when initial cross-evaluations diverge, allowing models to reconcile conflicting viewpoints and yield a robust, consensus-driven correctness score. We then integrate this estimate via an adaptive weighted loss function, enabling the student model to internalize the reliability of the training data. Extensive experiments on commonsense reasoning (CQA) and natural language inference (ANLI) benchmarks demonstrate that our method effectively improves student calibration while maintaining prediction accuracy.

Our contributions are:

1. Through a systematic analysis of student calibration across diverse distillation settings, we identify the absence of uncertainty information as the primary cause of miscalibration in black-box paradigms.
2. We propose an uncertainty-integrated framework for black-box knowledge distillation, designed to improve student calibration despite the inaccessibility of teacher confidence.
3. Empirical evaluations on the standard benchmarks CQA and ANLI, demonstrate that UIKD significantly enhances student calibration without compromising predictive accuracy. These findings offer a novel perspective for constructing trustworthy and reliable distilled models.

2 Related Work

Knowledge Distillation (KD) involves extracting knowledge from a teacher model to guide student model training, thereby enabling knowledge transfer and compression (Fang et al., 2026; Wang et al.,

2023a; Ho et al., 2023; Xu et al., 2024). Existing research on KD predominantly focuses on improving the accuracy of student models. Depending on teacher accessibility, KD can be categorized into white-box and black-box settings. In the white-box setting, logit-based KD transfers knowledge by matching the student’s output logits to the teacher’s. Building on this, some studies have examined how the teacher model’s confidence affects student accuracy. For example, Zhao et al. (2022) analyzes, both theoretically and experimentally, the impact of the teacher’s non-target class confidences, while Fan et al. (2024) shows a positive correlation between the teacher’s calibration and student accuracy. In the black-box setting, recent work (Hsieh et al., 2023) has instead emphasized using LLM-generated chain-of-thought (CoT) reasoning to boost student performance.

In contrast, research addressing the calibration of student models remains scarce. Shum et al. (2024) attempt to mitigate student miscalibration by applying temperature scaling to the teacher’s logits. However, this approach is constrained by two critical dependencies: it requires ground-truth labels for teacher calibration (which are often absent in KD scenarios) and mandates access to internal logits, rendering it inapplicable to black-box models. Moreover, the field currently lacks a systematic investigation into the underlying factors driving student calibration.

Model Calibration seeks to align a model’s confidence scores with the accuracy of its predictions. In black-box scenarios, where internal parameters are inaccessible, confidence is estimated solely based on inputs and outputs. The primary methods include: (1) Language-based confidence (Tian et al., 2023), which prompts the model to verbally express its certainty; and (2) Consistency-based estimation (Wang et al., 2023b), which measures con-

fidence by analyzing coherence across multiple outputs. While effective, the latter approach is inherently limited to the knowledge contained within a single model.

Conversely, for white-box models where internal parameters are accessible, calibration is typically conducted during either the alignment or inference phases. Alignment-phase methods include: (1) Label smoothing (Desai and Durrett, 2020), which applies a smoothing factor to convert hard labels into soft targets during training; and (2) Uncertainty elicitation (Xiong et al., 2024), which trains the model to explicitly generate uncertainty estimates alongside its answers. During the inference phase, the dominant approach is temperature scaling (Jiang et al., 2021; Guo et al., 2017), which modulates the temperature parameter to recalibrate output probabilities. However, this method requires an additional validation set for optimization. Furthermore, these strategies fail to account for data uncertainty, which is crucial for model calibration, thereby reducing their effectiveness in knowledge distillation with noisy data.

3 Preliminaries

Knowledge Distillation leverages a teacher model to generate pseudo-labels for task $\mathcal{I}$, constructing the dataset $D_{\mathcal{I}}^{\text{KD}}$. The student model is subsequently trained to emulate the teacher by minimizing the loss function $\mathcal{L}$. This procedure is formalized as follows:

$$D_{\mathcal{I}}^{\text{KD}} = \{\text{Parse}(o, s) \mid o \sim p_T(o \mid \mathcal{I}, s)\} \quad (1)$$

$$\mathcal{L} = \sum_{\mathcal{I}} \mathcal{L}_{\mathcal{I}}(D_{\mathcal{I}}^{\text{KD}}; \theta), \quad (2)$$

where $\text{Parse}(o, s)$ denotes a KD data instance extracted from the teacher model’s output o , s denotes examples of the seed knowledge, p_T denotes the teacher model, and θ represents the internal parameters of the student model.

Calibration A well-calibrated model assigns confidence scores that correctly reflect the true probability of its predictions being correct:

$$\mathbb{P}(Y = \hat{Y} \mid \text{conf}(X, \hat{Y}) = p) = p, \quad \forall p \in [0, 1]. \quad (3)$$

where X denotes the model input, Y the ground-truth label, $\hat{Y}$ the predicted label, $\text{conf}(\cdot, \cdot)$ the confidence function, and p a probability value.

Model calibration is evaluated using the *expected calibration error* (ECE). ECE is calculated as:

$$\text{ECE} = \sum_{b=1}^{N_b} \frac{|I_b|}{N_b} |\text{conf}(b) - \text{acc}(b)|, \quad (4)$$

where N_b denotes the number of bins; I_b and $|I_b|$ represent the samples and their count in bin b ; $\text{conf}(b)$ and $\text{acc}(b)$ denote the mean confidence and empirical accuracy, respectively.

Depending on the bucketing strategy, ECE is distinguished as either $\text{ECE}_{\text{interval}}$ or ECE_{mass} . $\text{ECE}_{\text{interval}}$ partitions the probability space into equal-width bins, whereas ECE_{mass} uses equal-frequency bins to handle uneven data distributions.

4 An Analysis of Calibration in KD

In this section, we conduct an empirical analysis to address a key question: *how does the knowledge distillation configuration influence the calibration of student models?*

Experimental Setup We conduct distillation experiments across a diverse set of configurations, varying teacher accessibility (white- vs. black-box), model scales, and architectures. To rigorously simulate black-box distillation, we use standard models but withhold their internal confidence scores during the process. Our evaluation spans two distinct architectural families: decoder-only models (LLaMA) (Team, 2024) and encoder-decoder models (T5) (Raffel et al., 2020). Within each series, we assess a range of teacher and student sizes on the CommonsenseQA (Talmor et al., 2019) and Adversarial NLI (Nie et al., 2020) benchmarks.

Results and Analysis Table 1 presents the results on the LLaMA family, in which the target distribution for black-box distillation is deterministic, so the confidence is set to 1 when computing the ECE. These results support two primary conclusions. First, student distilled from a black-box teacher generally exhibit poor calibration. For example, when distilling from LLaMA-8B to LLaMA-1B, the black-box setting produces an ECE of 41.86%, whereas white-box distillation achieves a substantially lower ECE of 28.71%. Although larger student tend to perform better, the disadvantage of black-box distillation persists across all model sizes, suggesting that the absence of teacher confidence detrimentally impacts student reliability. Second, the quality of the distillation data is critical; higher accuracy in the distillation data correlates with lower student ECE and improved calibration.

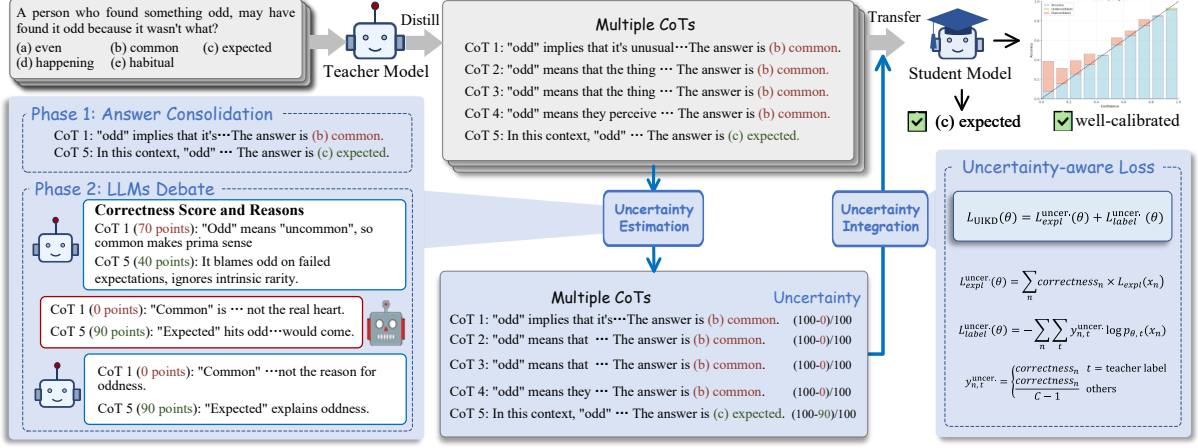


Figure 2: Overview of UIKD, which integrates two components: (1) Uncertainty Estimation of Knowledge Distillation Data and (2) Uncertainty-aware Loss Mechanism.

Type	ANLI				CQA			
	Teacher		Student		Teacher		Student	
	acc $\uparrow$	ECE $_{interval}\downarrow$	acc $\uparrow$	ECE $_{interval}\downarrow$	acc $\uparrow$	ECE $_{interval}\downarrow$	acc $\uparrow$	ECE $_{interval}\downarrow$
LLaMA3.1-8B$\rightarrow$LLaMA3.2-1B								
White-box	78.65	11.18	55.00	28.71	81.23	18.40	73.46	17.42
Black-box	78.65	21.35	55.30	41.86	81.23	19.77	71.58	25.82
LLaMA3.1-8B$\rightarrow$LLaMA3.2-3B								
White-box	78.65	11.18	63.50	20.78	81.23	18.40	78.71	13.82
Black-box	78.65	21.35	62.50	34.60	81.23	19.77	77.72	20.56
LLaMA3.2-3B$\rightarrow$LLaMA3.2-1B								
White-box	70.79	2.07	45.40	28.88	72.63	14.20	70.60	19.78
Black-box	70.79	29.21	46.70	53.12	72.63	17.37	69.45	28.43

Table 1: Accuracy and calibration (ECE $_{interval}$) of teacher and student models under different knowledge distillation settings on ANLI and CQA. For black-box KD, teacher-side ECE is computed using the exposed deterministic pseudo-labels with confidence set to 1.

The same behavior is observed for the T5 family, experimental details can be found in Appendix A.

Analysis of the Calibration Gap In the white-box setting, the loss function for a sample x_n , denoted by $\ell_{WB}(x_n; \theta)$, can be expressed as

$$\ell_{WB}(x_n; \theta) = - \sum_{i=1}^C p_i^*(x_n) \sum_{j=1}^C T_{ij} \log p_{\theta,j}(x_n). \quad (5)$$

where θ denotes the internal parameters of the student model, $p^*(x) = (P(Y = 1 | x), \dots, P(Y = C | x))$ denotes the true posterior distribution, C is the number of classes, and $T_{ij} = P(\hat{Y} = j | Y = i)$ represents an element of the transition matrix T .

In contrast, in the black-box setting, the corresponding loss function $\ell_{BB}(x_n; \theta)$ is given by

$$\ell_{BB}(x_n; \theta) = - \log p_{\theta, \hat{y}(x_n)}(x_n). \quad (6)$$

Based on this formulation, we have

$$\mathbb{E} [\|g_{BB}\|_2^2 | Y = i, X = x] = \|g_{WB}(x, i)\|_2^2 + 1 - \|T_i\|_2^2. \quad (7)$$

Here, g_{BB} and g_{WB} denote the gradients of the black-box and white-box loss functions, respectively. The detailed theoretical analysis is provided in Appendix B.

Implication for Method Design The analysis above shows that once black-box teacher outputs are collapsed into one-hot pseudo-labels, instance-level reliability is no longer encoded in the supervision signal. As a result, the student cannot distinguish between reliable and ambiguous distillation instances, leading to an increase in the variance of the loss gradient. This motivates a method that first estimates instance-level reliability from black-box outputs and then incorporates it into the training objective.

5 Method

Motivated by the analysis in Section 4, we propose Uncertainty-Integrated Knowledge Distillation (UIKD), a framework designed to enhance student model calibration. As illustrated by the blue part in Figure 2, UIKD augments the standard distillation process with two complementary components: uncertainty estimation of KD data and an uncertainty-aware loss mechanism.

To facilitate the subsequent uncertainty estimation and distillation processes, we first generate teacher knowledge. Following the approach of

Liu et al. (2024), we prompt the LLM to produce multiple CoTs alongside their corresponding answers. This method allows for systematic knowledge extraction and enables the verification of answer correctness. Additionally, we employ few-shot prompts (cf. Appendix H) to guide the teacher model toward generating higher-quality CoTs.

5.1 Uncertainty Estimation of KD Data

To compensate for the lack of explicit uncertainty signals in black-box knowledge distillation, we estimate the quality of the distillation data as a proxy for uncertainty. We employ a multi-LLM collaborative approach to objectively assess output correctness by leveraging the models’ internal knowledge. This method consists of two steps: *answer consolidation* and *LLMs debate*.

Answer Consolidation We observed that the teacher model frequently generates redundant answers across multiple reasoning paths. To prevent this redundancy from biasing the evaluation and to minimize token usage, we consolidate duplicate responses. For instance, if five generation sets yield only two unique results, the debate phase focuses exclusively on these distinct cases.

LLMs Debate To optimize token costs, we configure a binary debate between the teacher model and an auxiliary model, which are completely different from each other. Initially, both models independently evaluate the reasoning results, providing correctness scores (CS) with justifications. If the score discrepancy exceeds a threshold $a_{\max}$, a structured debate phase is triggered (Estornell and Liu, 2024). In each round of debate at this stage, the model is provided with a fixed debate template with the corresponding prompts, along with the ratings and justifications from both itself and the counterpart in the previous round. It then synthesizes its prior stance and the counterpart’s feedback, subsequently updating its correctness score (CS) while revising its justifications to reflect these diverse perspectives (cf. left part of Figure 2). The debate terminates after m rounds or when the score difference between agents falls below $a_{\max}$. The final CS $correctness_n$ of sample x_n is calculated as the average of the two agents’ ratings and is subsequently propagated to all reasoning paths associated with that answer. We provide the debate prompt in Appendix G.

5.2 Uncertainty-aware Loss Mechanism

To enable the student model to account for uncertainty within the distillation data, we incorporate uncertainty metrics directly into the training objective. Specifically, the loss function $\mathcal{L}_{\text{UIKD}}(\theta)$ comprises two components: explanation loss $\mathcal{L}_{\text{expl}}^{\text{uncer.}}(\theta)$ and answer loss $\mathcal{L}_{\text{label}}^{\text{uncer.}}(\theta)$.

$$\mathcal{L}_{\text{UIKD}}(\theta) = \mathcal{L}_{\text{expl}}^{\text{uncer.}}(\theta) + \mathcal{L}_{\text{label}}^{\text{uncer.}}(\theta) \quad (8)$$

For the explanation loss $\mathcal{L}_{\text{expl}}^{\text{uncer.}}(\theta)$, we apply the correctness of sample x_n (denoted as $correctness_n$) as a weighting factor to the explanation component $\mathcal{L}_{\text{expl}}$ of the standard black-box distillation loss:

$$\mathcal{L}_{\text{expl}}^{\text{uncer.}}(\theta) = \sum_n \left(correctness_n \times \mathcal{L}_{\text{expl}}(x_n) \right) \quad (9)$$

When calculating the loss for the answer, we assign the correctness score as the target probability for the correct token. The residual probability mass is then distributed uniformly across the remaining tokens:

$$y_{n,t}^{\text{uncer.}} = \begin{cases} correctness_n, & t = \text{teacher label}, \\ \frac{correctness_n}{C-1}, & \text{otherwise.} \end{cases} \quad (10)$$

$$\mathcal{L}_{\text{label}}^{\text{uncer.}}(\theta) = - \sum_n \sum_t y_{n,t}^{\text{uncer.}} \log p_{\theta,t}(x_n) \quad (11)$$

where $y_{n,t}^{\text{uncer.}}$ represents the label weight, C is the number of candidate labels.

6 Experiments

To evaluate the effectiveness of UIKD, we investigate the following research questions:

- **RQ1:** To what extent does UIKD reduce calibration error in student models?
- **RQ2:** How does UIKD impact the trade-off between model accuracy and calibration?
- **RQ3:** Is the efficacy of UIKD consistent across different model architectures and scales?

6.1 Experimental Setup

Datasets. We evaluate our proposed method on two distinct tasks: Commonsense Question Answering (CQA) using the CommonsenseQA dataset (Talmor et al., 2019), and Natural Language Inference (NLI) using the ANLI dataset (Nie et al., 2020). As established benchmarks in the knowledge distillation

Model	Scenarios	Method	CQA				ANLI			
			acc $\uparrow$	ECE $_{mass}\downarrow$	ECE $_{interval}\downarrow$	trust $\uparrow$	acc $\uparrow$	ECE $_{mass}\downarrow$	ECE $_{interval}\downarrow$	trust $\uparrow$
T5-base	answer-only	5 CoTs	59.71	24.59	24.33	35.12	48.40	40.94	40.90	7.46
		UE	58.31	38.89	38.96	19.42	<u>49.10</u>	37.08	35.73	12.02
		LS	59.05	19.75	19.71	39.30	44.90	44.33	44.33	0.57
		SC	61.49	<u>18.58</u>	18.83	<u>43.91</u>	48.60	36.81	35.42	11.79
		LC	<u>62.33</u>	19.53	<u>17.82</u>	<u>42.80</u>	48.50	36.99	35.74	11.51
		UIKD _{SC}	61.83	20.07	19.43	41.76	49.30	<u>35.87</u>	<u>34.15</u>	<u>13.43</u>
		UIKD _{LC}	<u>62.33</u>	19.82	18.45	42.51	48.70	36.46	34.66	12.24
		UIKD	62.90	16.28	15.68	46.62	49.30	34.22	33.58	15.08
	CoT	1 CoT	61.75	28.89	28.88	32.86	48.70	41.98	41.43	6.72
		5 CoTs	64.29	21.28	21.06	43.01	48.60	37.26	37.27	11.34
		SC	63.14	18.90	17.79	44.24	46.30	39.83	38.61	6.47
		LC	63.47	19.39	18.60	44.17	48.70	36.84	35.09	11.86
		UIKD _{SC}	64.45	18.80	<u>17.54</u>	45.65	46.80	37.44	36.43	9.36
		UIKD _{LC}	<u>64.37</u>	<u>18.63</u>	18.34	<u>45.74</u>	<u>49.00</u>	<u>32.14</u>	<u>30.31</u>	<u>17.86</u>
		UIKD	<u>64.37</u>	16.37	14.75	48.00	49.10	30.96	29.63	18.14
LLaMA-1B	answer-only	5 CoTs	62.24	19.24	18.19	43.01	59.80	21.15	19.62	38.65
		UE	63.13	28.66	28.76	34.47	60.50	40.18	32.08	20.32
		LS	63.23	13.99	13.49	49.24	61.10	18.48	<u>18.58</u>	<u>42.62</u>
		SC	62.98	12.95	9.88	50.03	60.90	20.94	20.79	39.96
		LC	63.37	13.67	12.47	49.70	61.30	22.47	22.78	38.83
		UIKD _{SC}	62.57	<u>11.73</u>	<u>7.59</u>	<u>50.84</u>	61.30	21.26	19.65	40.04
		UIKD _{LC}	63.88	15.81	15.28	48.07	<u>62.40</u>	22.35	22.10	40.05
		UIKD	<u>63.39</u>	10.78	7.23	52.61	62.70	<u>19.87</u>	17.96	42.83
	CoT	1 CoT	62.08	30.40	30.40	31.68	59.20	35.32	35.01	23.88
		5 CoTs	<u>64.70</u>	<u>17.40</u>	<u>17.40</u>	<u>47.30</u>	60.30	24.62	21.61	35.68
		SC	63.10	22.79	22.17	40.31	59.60	21.90	19.19	37.70
		LC	62.38	21.43	21.06	40.95	60.80	21.47	19.73	39.33
		UIKD _{SC}	63.80	22.98	20.37	40.82	<u>60.90</u>	18.74	<u>15.89</u>	42.16
		UIKD _{LC}	62.57	19.83	19.20	42.74	60.60	<u>16.75</u>	20.30	<u>43.85</u>
		UIKD	64.78	16.45	13.77	48.34	61.00	13.85	14.09	47.15

Table 2: Main Results: we compare the performance of various methods on the CQA and ANLI datasets using the T5-base and LLaMA-1B. Best results are in **bold**, and second-best results are underlined.

literature, these datasets provide a robust baseline for comprehensive evaluation. Dataset statistics are given in Appendix C.

Implementation Details. We select GPT-3.5-turbo as the teacher model. For student models, we employ architectures representing two distinct paradigms: *encoder-decoder* models from the T5 family (Raffel et al., 2020) and *decoder-only* models from the LLaMA family (Team, 2024). Our experiments are conducted under two KD settings: *answer-only distillation* (Iliopoulos et al., 2022) and *chain-of-thought distillation* (Hsieh et al., 2023). To enhance data quality and ensure consistent formatting during the generation of distillation data, we utilize 5-shot CoT prompts. Following the methodology of Liu et al. (2024), we sample 5 CoT paths for each training instance, effectively balancing the diversity of the distillation data against computational costs. During the UIKD debate phase, we leverage DeepSeek-V3 (DeepSeek-AI, 2024)

as an auxiliary agent to engage in debate with the teacher model. More implementation details are presented in Appendix D.

We employ acc (%) to measure the prediction accuracy of the student model and ECE (%) to assess its calibration, as detailed in Section 3. Higher acc indicates better predictive performance, while lower ECE indicates better calibration, with both reported as percentages. Furthermore, we adopt the trust metric to jointly evaluate predictive accuracy and calibration (Shum et al., 2024), defined as: $\text{trust} = \text{acc} - \text{ECE}_{\text{mass}}$.

Baselines. We compare our method against baselines categorized into two groups: (1) *KD baselines*. We consider two standard settings: answer-only distillation and CoT-based distillation. For each, we evaluate variants using either one or five teacher outputs. For CoT-based distillation, we also compare against other baseline methods based on multi-task learning (Li et al., 2024). (2) *Training-*

stage white-box calibration methods. We compare against Label Smoothing (LS) (Desai and Durrett, 2020), applying a smoothing coefficient of 0.1 during distillation, and Uncertainty Elicitation (UE) (Xiong et al., 2024), a recent SoTA approach that trains the model to explicitly output an uncertainty score alongside its answer. (3) Filtering-based method: We filter samples and perform reweighting based on the scores obtained from Self-Consistency (SC) and Language Calibration (LC) methods. Furthermore, we investigate two variants of our proposed method, UIKD_{SC} and UIKD_{LC} . These differ in their uncertainty estimation modules, adopting Self-Consistency (SC) (Wang et al., 2023b) and Language-Calibration (LC) (Tian et al., 2023), respectively. The implementation details of the baselines are provided in Appendix E.

6.2 Main Results

To what extent does UIKD reduce calibration error in student models? Table 2 details the calibration performance across various knowledge distillation methods. We observe that UIKD consistently surpasses the baselines in diverse scenarios, achieving lower ECE values that signify improved calibration. Notably, for the LLaMA-1B model on the ANLI dataset (under CoT distillation) ECE_{mass} and $\text{ECE}_{\text{interval}}$ decrease by 10.77% and 5.72%, respectively. These results substantiate the finding that incorporating data uncertainty into the KD process enhances student model calibration.

How does UIKD impact the trade-off between model accuracy and calibration? Table 2 demonstrates that UIKD consistently enhances student model accuracy alongside calibration. Notably, on the CQA dataset (under answer-only distillation), UIKD achieves a 2.71% improvement in the trust metric and a 0.57% accuracy boost, while simultaneously reducing ECE_{mass} by 2.30%. This confirms that superior calibration need not be achieved at the expense of predictive performance. In contrast, while baselines such as UE and LS offer calibration gains, they often incur accuracy penalties, showing reductions of 1.40% and 0.66% respectively in the same scenario. This further indicates that UIKD effectively rectifies confidence misalignment, mitigating cases where models are overconfident in erroneous predictions or underconfident in correct ones.

Is the efficacy of UIKD consistent across different model architectures and scales? Table 2 demon-

strates the consistent efficacy of UIKD across diverse architectures, yielding comparable ECE reductions for both decoder-only (LLaMA) and encoder-decoder (T5) models. Furthermore, Table 3 details the performance across varying model scales. The results indicate that UIKD achieves lower ECE values across the spectrum of student sizes, confirming its ability to generalize effectively. Notably, in the answer-only setting, UIKD surpasses baselines on T5-small, T5-base, and T5-large with improvements of 2.04%, 2.71%, and 1.70% in the trust metric, respectively. Collectively, these findings underscore the robustness of UIKD across disparate model architectures and scales.

6.3 Ablation Study

To isolate the specific contributions of our framework’s components, we progressively ablate the debate mechanism (w/o debate) and the uncertainty integration in the answer (w/o $\mathcal{L}_{\text{label}}^{\text{uncer.}}$) and explanation (w/o $\mathcal{L}_{\text{expl}}^{\text{uncer.}}$) losses. As detailed in Table 4, we observe that retaining uncertainty only in the answer loss (w/o $\mathcal{L}_{\text{expl}}^{\text{uncer.}}$) yields significant gains in both accuracy and calibration for the 5 CoTs, occasionally surpassing the full UIKD model in accuracy. In contrast, applying uncertainty solely to the explanation loss (w/o $\mathcal{L}_{\text{label}}^{\text{uncer.}}$) enhances calibration but comes at the expense of accuracy. These findings suggest that simultaneous incorporation of uncertainty in both answer and explanation stages is essential for UIKD to optimally balance reliability and performance. Finally, removing the debate mechanism (substituting it with LC for uncertainty estimation) leads to inferior calibration and accuracy, attributable to the inaccurate uncertainty estimation inherent in the LC baseline compared to our debate-driven approach.

6.4 Analysis

UIKD Improves Uncertainty Calibration via Multi-LLM Debate. UIKD consistently yields superior calibration compared to its variants, UIKD_{SC} and UIKD_{LC} . To elucidate the source of this improvement, we evaluate the quality of the uncertainty estimation on the distillation data, as detailed in Table 5. We observe that the Multi-LLM debate mechanism achieves the lowest error rates among the three methods, with $\text{ECE}_{\text{mass}}/\text{ECE}_{\text{interval}}$ values of 12.41%/18.66% on CQA and 8.53%/13.91% on ANLI. This confirms that leveraging the collective internal knowledge of multiple LLMs enables

Scenarios	Method	T5-small				T5-base				T5-large			
		acc $\uparrow$	ECE $_{mass}\downarrow$	ECE $_{interval}\downarrow$	trust $\uparrow$	acc $\uparrow$	ECE $_{mass}\downarrow$	ECE $_{interval}\downarrow$	trust $\uparrow$	acc $\uparrow$	ECE $_{mass}\downarrow$	ECE $_{interval}\downarrow$	trust $\uparrow$
answer-only	5 CoTs	46.27	22.51	20.88	23.76	59.71	24.59	24.33	35.12	64.95	20.08	19.51	44.87
	UE	43.65	55.86	55.86	-12.21	58.31	38.89	38.96	19.42	64.13	28.66	28.76	35.47
	LS	45.21	23.06	21.37	22.15	59.05	19.75	19.71	39.30	68.71	13.28	11.24	55.43
	SC	45.29	20.73	<u>18.18</u>	24.56	61.49	<u>18.58</u>	18.83	<u>43.91</u>	67.35	15.68	13.49	51.67
	LC	46.11	22.24	21.48	23.87	<u>62.33</u>	19.53	<u>17.82</u>	42.80	68.97	17.09	14.87	51.88
	UIKD _{SC}	45.78	<u>19.98</u>	18.49	<u>25.80</u>	<u>61.83</u>	20.07	19.43	41.76	67.32	14.18	11.83	53.14
	UIKD _{LC}	46.36	21.08	18.67	25.28	<u>62.33</u>	19.82	18.45	42.51	<u>69.04</u>	16.34	15.28	52.70
	UIKD	46.60	18.76	15.46	27.84	62.90	16.28	15.68	46.62	70.02	12.89	10.14	57.13
CoT	1 CoT	46.60	32.50	31.89	14.10	61.75	28.89	28.88	32.86	<u>70.76</u>	25.63	25.67	45.13
	5 CoTs	45.13	25.18	23.60	19.95	64.29	21.28	21.06	43.01	69.94	16.57	15.21	53.37
	SC	45.78	23.21	20.53	22.57	63.14	18.90	17.79	44.24	70.18	12.42	10.62	57.76
	LC	46.27	23.44	21.07	22.83	63.47	19.39	18.60	44.17	69.80	<u>11.69</u>	10.47	58.11
	UIKD _{SC}	47.91	<u>19.94</u>	<u>17.48</u>	<u>27.97</u>	64.45	18.80	<u>17.54</u>	45.65	70.52	12.32	<u>10.42</u>	<u>58.20</u>
	UIKD _{LC}	48.89	21.30	18.44	27.59	<u>64.37</u>	<u>18.63</u>	18.34	<u>45.74</u>	69.86	12.00	10.86	57.86
	UIKD	<u>48.16</u>	19.41	17.47	28.75	<u>64.37</u>	16.37	14.75	48.00	71.58	10.91	7.30	60.67

Table 3: Experimental results on models of different scales. Best results are in **bold**; second-best results are underlined.

Model	Method	CQA				ANLI			
		acc $\uparrow$	ECE $_{mass}\downarrow$	ECE $_{interval}\downarrow$	trust $\uparrow$	acc $\uparrow$	ECE $_{mass}\downarrow$	ECE $_{interval}\downarrow$	trust $\uparrow$
T5-base	5 CoTs	64.29	21.28	21.06	43.01	48.60	37.26	37.27	11.34
	UIKD	64.37	16.37	14.75	48.00	49.10	30.96	29.63	18.14
	w/o $\mathcal{L}_{\text{factual}}$	63.65	20.04	20.22	43.61	48.49	35.91	36.80	12.58
	w/o $\mathcal{L}_{\text{epist}}$	64.70	18.44	17.74	46.26	49.90	33.87	32.15	16.03
	w/o debate	62.33	19.82	18.45	42.51	49.00	32.46	30.31	16.54
LLaMA	5 CoTs	64.70	17.40	17.40	47.30	60.30	24.62	21.61	35.68
	UIKD	64.78	16.45	13.77	48.33	61.00	13.85	14.09	47.15
	w/o $\mathcal{L}_{\text{factual}}$	62.84	17.01	17.48	45.83	60.00	20.80	22.82	39.20
	w/o $\mathcal{L}_{\text{epist}}$	64.80	16.66	16.62	48.16	61.00	16.92	17.77	44.08
	w/o debate	62.57	19.83	19.20	42.74	60.60	16.75	20.30	43.85

Table 4: Ablation study on UIKD method.

a more accurate assessment of data uncertainty, which in turn drives the improved calibration of the student model.

Method	CQA		ANLI	
	ECE $_{mass}\downarrow$	ECE $_{interval}\downarrow$	ECE $_{mass}\downarrow$	ECE $_{interval}\downarrow$
black-box	27.71	27.71	16.71	16.71
SC	13.52	22.97	9.88	14.58
LC	18.11	20.72	13.90	19.43
UIKD	12.41	18.66	8.53	13.91

Table 5: Performance of Different Uncertainty Estimation Methods. Best results are in **bold**.

UIKD yields more reliable student under selective generation. Selective generation serves as a critical downstream application of calibration (Lee et al., 2024), requiring a model to abstain from answering when its uncertainty is high. We analyze the performance of student models in this setting to benchmark diverse calibration methods, as presented in Table 6. At identical confidence thresholds, UIKD and its variants consistently yield higher accuracy than the baselines. Notably, UIKD (leveraging Multi-LLM debate) achieves the highest accuracy across diverse thresholds, outperforming baselines by up to 6.44%. This result further validates the practical utility and robustness of UIKD in real-

world scenarios.

UIKD improves the reliability of the student model. Reliability diagrams (Xia et al., 2025b) provide an intuitive measure of the alignment between predicted confidence and empirical accuracy. As shown in Figure 4 (Appendix I), we compare these profiles across varying distillation methods. The results demonstrate that UIKD achieves superior calibration: unlike baselines, it effectively mitigates both underconfidence and overconfidence across all confidence intervals.

Threshold	5 CoTs		LS		UIKD _{LC}		UIKD _{SC}		UIKD	
	acc	num $_{\text{sample}}$	acc	num $_{\text{sample}}$	acc	num $_{\text{sample}}$	acc	num $_{\text{sample}}$	acc	num $_{\text{sample}}$
0.5	62.33	1115	62.37	1079	64.49	1073	67.24	980	67.97	1021
0.6	64.11	1031	64.57	988	67.18	975	69.07	902	70.55	927
0.7	66.88	945	67.05	868	69.46	871	70.81	812	72.79	827
0.8	70.19	842	69.81	732	73.23	747	72.36	720	75.07	710
0.9	74.48	678	75.58	520	77.51	587	77.36	552	80.74	488

Table 6: Accuracy of calibrated models on accepted predictions under different confidence thresholds in the selective generation setting. num $_{\text{sample}}$ denotes the number of samples whose confidence scores exceed a given threshold. Best results are in **bold**.

a_{max}	m	acc $\uparrow$	ECE $_{mass}\downarrow$	trust $\uparrow$	avg $_{\text{call}}$
5	2	64.16	17.31	46.85	2.57
5	4	64.37	16.37	48.00	2.75
5	6	64.39	16.35	48.04	2.79
10	4	64.01	18.01	46.00	2.48

Table 7: Sensitivity analysis of UIKD with respect to the debate threshold a_{max} and the maximum number of debate rounds m on CQA.

Sensitivity to Debate Hyperparameters. We further analyze the sensitivity of UIKD to a_{max} and m on CQA. The results in Table 7 show that UIKD is stable around the default setting (a_{max}, m) =

(5, 4). Increasing m from 4 to 6 yields only a 0.04 improvement in Trust, while increasing avg_call from 2.75 to 2.79, which may lead to higher token consumption over the full dataset. Therefore, $(a_{\max}, m) = (5, 4)$ provides a good trade-off between performance and cost.

Case Study. We conduct case studies under the answer-only distillation setting to qualitatively evaluate model performance, with detailed results presented in Appendix J. Across examples from CQA and ANLI, UIKD produces confidence scores that better reflect prediction correctness. In particular, it suppresses overconfidence when predictions are incorrect and assigns high confidence to correct predictions. In addition, we further analyze a failure mode of debate-based uncertainty estimation, where both debate agents assign high correctness scores to the same incorrect answer; see Appendix K.

Computational Overhead of UIKD. To quantify the potential overhead introduced during the multi-LLM debate stage, we measure and compare the generated token consumption-per-sample on the distillation dataset across different methods, as reported in Table 8. We also report the average API calls per example (avg_call) and the debate trigger rate during the debate stage in Table 9. The results indicate that the multi-LLM debate mechanism introduces only a modest increase in token usage while remaining computationally manageable.

Method	CQA	ANLI
CoT distill	83.55	101.36
Label Smoothing	83.55	101.36
Uncertainty Elicitation	84.55	102.36
UIKD	91.37	110.67

Table 8: Computational Overhead of UIKD on CQA and ANLI datasets.

Dataset	avg_call	debate_trigger_rate
CQA	2.75	42.85
ANLI	2.97	35.88

Table 9: Average API calls per example and debate trigger rate during the debate stage.

7 Conclusion

This paper presents a systematic evaluation of student model calibration across diverse knowledge distillation settings. We identify the absence of data

uncertainty as the primary factor driving miscalibration in black-box scenarios. To remedy this, we propose **UIKD**, a framework specifically designed to restore calibration in black-box settings by synthesizing uncertainty signals. Ultimately, this work establishes a new foundation for constructing trustworthy and reliable student models.

Limitations

This work investigates student model calibration and proposes UIKD to enhance it. The experimental results demonstrate the effectiveness of the proposed approach, and there are clear opportunities for further extension and refinement. On one hand, due to practical cost considerations, the experiments primarily adopt GPT-3.5-turbo as the teacher model, and evaluating a wider variety of teacher models would further strengthen evidence for the generality of the proposed approach. On the other hand, although the method is mainly designed for classification settings, generation-task experiments are also included in Appendix F, and extending UIKD more fully to generation tasks may require additional design refinements beyond the scope of this work.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.62576200) and the Fundamental Research Program of Shanxi Province (No.202503021211076).

References

- Yao Chen, Jiawei Sheng, Wenyuan Zhang, and Tingwen Liu. 2025. [Improving reasoning capabilities in small models through mixture-of-layers distillation with stepwise attention on key information](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4952–4971, Suzhou, China. Association for Computational Linguistics.
- DeepSeek-AI. 2024. [Deepseek-v3 technical report](#). *CoRR*, abs/2412.19437.
- DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025. [Deepseek-v3.2: Pushing the frontier of open large language models](#). *Preprint*, arXiv:2512.02556.

- Shrey Desai and Greg Durrett. 2020. [Calibration of pre-trained transformers](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 295–302. Association for Computational Linguistics.
- Andrew Estornell and Yang Liu. 2024. [Multi-llm debate: Framework, principals, and interventions](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 28938–28964. Curran Associates, Inc.
- Wen-Shu Fan, Su Lu, Xin-Chun Li, De-Chuan Zhan, and Le Gan. 2024. [Revisit the essence of distilling knowledge through calibration](#). In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pages 12882–12894. PMLR / OpenReview.net.
- Luyang Fang, Xiaowei Yu, Jiazhang Cai, Yongkai Chen, Shushan Wu, Zhengliang Liu, Zhenyuan Yang, Hao-ran Lu, Xilin Gong, Yufang Liu, Terry Ma, Wei Ruan, Ali Abbasi, Jing Zhang, Tao Wang, Ehsan Latif, Wei Liu, Wei Zhang, Soheil Kolouri, and 5 others. 2026. [Knowledge distillation and dataset distillation of large language models: emerging trends, challenges, and future directions](#). *Artif. Intell. Rev.*, 59(1):17.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. [On calibration of modern neural networks](#). In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.
- Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. [Large language models are reasoning teachers](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14852–14882, Toronto, Canada. Association for Computational Linguistics.
- Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. [Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, Toronto, Canada. Association for Computational Linguistics.
- Fotis Iliopoulos, Vasilis Kontonis, Cenk Baykal, Gaurav Menghani, Khoa Trinh, and Erik Vee. 2022. [Weighted distillation with unlabeled examples](#). In *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*.
- Wangyi Jiang, Yaojie Lu, Hongyu Lin, Xianpei Han, and Le Sun. 2025. [Teach small models to reason by curriculum distillation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7423–7433, Suzhou, China. Association for Computational Linguistics.
- Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. [How can we know When language models know? on the calibration of language models for question answering](#). *Trans. Assoc. Comput. Linguistics*, 9:962–977.
- Minjae Lee, Kyungmin Kim, Taesoo Kim, and Sangdon Park. 2024. [Selective generation for controllable language models](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Shiyang Li, Jianshu Chen, yelong shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, Wenhui Chen, and Xifeng Yan. 2024. [Explanations from large language models make small reasoners better](#). In *2nd Workshop on Sustainable AI*.
- Weize Liu, Guocong Li, Kai Zhang, Bang Du, Qiyuan Chen, Xuming Hu, Hongxia Xu, Jintai Chen, and Jian Wu. 2024. [Mind’s mirror: Distilling self-evaluation capability and comprehensive thinking from large language models](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 6748–6763. Association for Computational Linguistics.
- Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. [Adversarial NLI: A new benchmark for natural language understanding](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.
- Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. [Are NLP models really able to solve simple math word problems?](#) In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online. Association for Computational Linguistics.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *J. Mach. Learn. Res.*, 21:140:1–140:67.
- KaShun Shum, Minrui Xu, Jianshu Zhang, Zixin Chen, Shizhe Diao, Hanze Dong, Jipeng Zhang, and Muhammad Omer Raza. 2024. [FIRST: Teach a reliable large language model through efficient trustworthy distillation](#). In *Proceedings of the 2024 Confer-*

- ence on Empirical Methods in Natural Language Processing, pages 12646–12659, Miami, Florida, USA. Association for Computational Linguistics.
- 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4149–4158. Association for Computational Linguistics.
- Llama Team. 2024. [The llama 3 herd of models](#). *CoRR*, abs/2407.21783.
- Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. [Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.
- Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. 2023a. [SCOTT: Self-consistent chain-of-thought distillation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5546–5558, Toronto, Canada. Association for Computational Linguistics.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. [Self-consistency improves chain of thought reasoning in language models](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Mingxuan Xia, Haobo Wang, Yixuan Li, Zewei Yu, Jindong Wang, Junbo Zhao, and Runze Wu. 2025a. [Prompt candidates, then distill: A teacher-student framework for LLM-driven data annotation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2750–2770, Vienna, Austria. Association for Computational Linguistics.
- Yuxi Xia, Pedro Henrique Luz de Araujo, Klim Zaporozhets, and Benjamin Roth. 2025b. [Influences on LLM calibration: A study of response agreement, loss functions, and prompt styles](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2025, Vienna, Austria, July 27 - August 1, 2025, pages 3740–3761. Association for Computational Linguistics.
- Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. [Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Hao Xu, Yunxiao Zhao, Jiayang Zhang, Zhiqiang Wang, and Ru Li. 2025. [LOG: A local-to-global optimization approach for retrieval-based explainable multi-hop question answering](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9085–9095, Abu Dhabi, UAE. Association for Computational Linguistics.
- Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. 2024. [A survey on knowledge distillation of large language models](#). *CoRR*, abs/2402.13116.
- Zhichao Yan, Jiapu Wang, Jiaoyan Chen, Yanyan Wang, Hongye Tan, Jiye Liang, Xiaoli Li, Ru Li, and Jeff Z Pan. 2025. Prompting large language models with partial knowledge for answering questions with unseen entities. *arXiv preprint arXiv:2508.01290*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 40 others. 2025a. [Qwen3 technical report](#). *CoRR*, abs/2505.09388.
- Chuanpeng Yang, Yao Zhu, Wang Lu, Yidong Wang, Qian Chen, Chenlong Gao, Bingjie Yan, and Yiqiang Chen. 2025b. [Survey on knowledge distillation for large language models: Methods, evaluation, and application](#). *ACM Trans. Intell. Syst. Technol.*, 16(6).
- Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. 2022. [Decoupled knowledge distillation](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 11943–11952. IEEE.

A Experimental Results on the T5 Model

Table 10 presents the experimental results of the T5 model on the ANLI and CQA datasets. The results indicate that the conclusions drawn from the T5 model are consistent with the conclusions drawn in Section 4 based on the LLaMA model.

Type	ANLI				CQA			
	Teacher		Student		Teacher		Student	
	acc $\uparrow$	ECE $_{interval}$ $\downarrow$	acc $\uparrow$	ECE $_{interval}$ $\downarrow$	acc $\uparrow$	ECE $_{interval}$ $\downarrow$	acc $\uparrow$	ECE $_{interval}$ $\downarrow$
T5-large$\rightarrow$T5-small								
White-box	93.02	1.97	42.20	40.11	75.52	4.82	44.31	12.44
Black-box	81.23	6.98	42.10	50.71	75.52	24.48	40.54	16.31
T5-large$\rightarrow$T5-base								
White-box	93.02	1.97	44.80	40.16	75.52	4.82	64.62	12.02
Black-box	81.23	6.98	45.30	45.47	75.52	24.48	59.71	12.18
T5-base$\rightarrow$T5-small								
White-box	79.98	2.98	41.00	43.59	71.29	3.36	39.89	18.36
Black-box	79.98	20.02	42.10	52.99	71.29	28.71	39.64	27.64

Table 10: Performance and calibration (ECE $_{interval}$) of teacher and student models under different knowledge distillation settings on ANLI and CQA. For black-box KD, teacher-side ECE is computed using the exposed deterministic pseudo-labels with confidence set to 1.

B Theoretical Analysis

In this section, we provide a theoretical analysis to explain the calibration gap observed between student models distilled in black-box versus white-box settings.

Let $\mathcal{D} = \{(x_k, y_k)\}_{k=1}^N$ be a dataset that follows the distribution $P(x, y)$, where x represents the input features and $y \in \{1, 2, \dots, C\}$ is the true class label. Due to the limitations of the knowledge within the teacher model, the knowledge distillation data generated by the teacher model may contain some errors. Assume that the data labels $\hat{y}$ contain noise, where the probability of noise occurrence follows a transition matrix T :

$$T_{ij} = P(\hat{Y} = j \mid Y = i), \quad \sum_{j=1}^C T_{ij} = 1. \quad (12)$$

Let the true posterior be $p^*(x) = (P(Y = 1 \mid x), \dots, P(Y = C \mid x))$. Then the observed teacher soft target is:

$$\hat{p}(x) = T^\top p^*(x). \quad (13)$$

Consequently, the white-box distillation loss $\mathcal{L}_{WB}$ should be written as:

$$\begin{aligned} \mathcal{L}_{WB}(\theta) &= -\mathbb{E}_x \sum_{j=1}^C \hat{p}_j(x) \log p_{\theta,j}(x) \\ &= -\mathbb{E}_x \sum_{i=1}^C p_i^*(x) \sum_{j=1}^C T_{ij} \log p_{\theta,j}(x), \end{aligned} \quad (14)$$

where θ denotes the trainable parameters of the student model.

If we further define:

$$\begin{aligned} \hat{y}(x) &= \arg \max_j \hat{p}_j(x) \\ A_n^{WB} &= \mathbf{1}\{\hat{y}(x_n) = y_n\} \end{aligned} \quad (15)$$

Then Eq. (14) can be computed as:

$$\begin{aligned} \hat{\mathcal{L}}_{WB}(\theta) &= \frac{1}{N} \sum_{n=1}^N A_n^{WB} \ell_{WB}(x_n; \theta) \\ &\quad + \frac{1}{N} \sum_{n=1}^N (1 - A_n^{WB}) \ell_{WB}(x_n; \theta). \end{aligned} \quad (16)$$

where

$$\ell_{WB}(x_n; \theta) = \sum_{i=1}^C p_i^*(x) \sum_{j=1}^C T_{ij} \log p_{\theta,j}(x). \quad (17)$$

The black-box loss function corresponding to the above white-box loss can then be expressed as:

$$\begin{aligned} \hat{\mathcal{L}}_{BB}(\theta) &= \frac{1}{N} \sum_{n=1}^N A_n [-\log p_{\theta, \hat{y}(x_n)}(x_n)] \\ &\quad + \frac{1}{N} \sum_{n=1}^N (1 - A_n) [-\log p_{\theta, \hat{y}(x_n)}(x_n)]. \end{aligned} \quad (18)$$

where $A_n = \mathbf{1}[\hat{y}(x_n) = y_n]$.

Comparing the two loss functions, we observe that data uncertainty is absent from the loss function due to the black-box nature of the teacher model.

Building on this, let $z_\theta(x) = (z_{\theta,1}(x), \dots, z_{\theta,C}(x))$ denote the logits output by the student model before the softmax. The white-box sample-level gradient $g_{WB}(x)$ with respect to the logits is given by

$$g_{WB}(x) = \nabla_{z_\theta(x)} \ell_{WB}(x; \theta) = p_\theta(x) - \hat{p}(x). \quad (19)$$

When the ground-truth label is y , we obtain

$$g_{WB}(x, y) = p_\theta(x) - T_y. \quad (20)$$

where T_y refers to the y -th row of the transition matrix T .

For black-box knowledge distillation, where the supervision signal degenerates to a one-hot pseudo-label $e_{\hat{y}}$, the corresponding sample-level gradient $g_{\text{BB}}(x, \hat{y})$ is given by:

$$g_{\text{BB}}(x, \hat{y}) = \nabla_{z_{\theta}(x)} \ell_{\text{BB}}(x, \hat{y}; \theta) = p_{\theta}(x) - e_{\hat{y}}. \quad (21)$$

The difference in squared gradient norms between $g_{\text{BB}}^{\text{inc}(j)}$, the gradient induced by an incorrect teacher top-1 label j , and $g_{\text{BB}}^{\text{cor}}$, the gradient induced by a correct one, can be expressed as:

$$\|g_{\text{BB}}^{\text{inc}(j)}\|_2^2 - \|g_{\text{BB}}^{\text{cor}}\|_2^2 = 2(p_y - p_j). \quad (22)$$

As long as the student still favors the true class over the incorrect class, i.e., $p_y > p_j$, we have

$$\|g_{\text{BB}}^{\text{inc}(j)}\|_2 > \|g_{\text{BB}}^{\text{cor}}\|_2. \quad (23)$$

This shows that once the student begins to learn the correct trend, an incorrect teacher top-1 produces a larger sample-level gradient than a correct teacher top-1.

Furthermore, the gradient variance can be computed as:

$$\begin{aligned} & \text{Var}_{\hat{Y}|Y=i}(g_{\text{BB}}) \\ &:= \mathbb{E} [|g_{\text{BB}} - \mathbb{E} g_{\text{BB}}|_2^2 | Y = i, X = x] \\ &= \mathbb{E} [|e_{\hat{Y}} - T_i|_2^2 | Y = i] \\ &= \mathbb{E} [|e_{\hat{Y}}|_2^2 | Y = i] - 2T_i^T \mathbb{E}[e_{\hat{Y}} | Y = i] + |T_i|_2^2 \\ &= 1 - |T_i|_2^2 \end{aligned} \quad (24)$$

Combined with the variance formula, we have

$$\mathbb{E} [|g_{\text{BB}}|_2^2 | Y = i, X = x] = |g_{\text{WB}}(x, i)|_2^2 + 1 - |T_i|_2^2. \quad (25)$$

The analysis above indicates that, due to the black-box nature of the teacher model, data uncertainty is absent from the loss function. This lack of differentiation, manifested as increased variance in the loss function gradient, exacerbates the negative impact on the student model.

This observation motivates us to explicitly estimate data uncertainty and incorporate it into the distillation process by reweighting the cross-entropy loss according to the reliability of each distillation instance.

C Dataset Statistics and Splits

Table 11 reports the dataset split statistics for CQA and ANLI. We follow the official splits provided by the original benchmarks to ensure fair comparison and reproducibility.

Dataset	Train	Validation	Test
CQA	8766	975	1221
ANLI	16946	1000	1000

Table 11: Statistics of dataset splits for CQA and ANLI.

D Experimental Details

All models are trained and evaluated on a single NVIDIA A100 GPU with 40GB memory. For the T5 family, we follow the training settings of Liu et al. (2024), including the optimizer configuration, batch size, random seed, and other hyperparameters, to ensure a fair comparison. For the LLaMA-based models, we train the models for 3 epochs with a learning rate of 1e-6. For each model, the final checkpoint is used for evaluation. During the experiments, we set the debate threshold a_{max} to 5 and the maximum number of debate rounds m to 4.

E Baselines

We further elaborate on the implementation details of the baselines as follows:

Label Smoothing We convert the answers generated by the teacher model into one-hot labels, and compute the knowledge distillation loss based on the label-smoothed one-hot labels. The loss function $\mathcal{L}_{\text{LS}}(\theta)$ is given by:

$$y_{n,t}^{\text{LS}} = \begin{cases} 1 - \epsilon + \frac{\epsilon}{C}, & t = y_n \\ \frac{\epsilon}{C}, & t \neq y_n \end{cases} \quad (26)$$

$$\mathcal{L}_{\text{LS}}(\theta) = - \sum_i \sum_t y_{n,t}^{\text{LS}} \log p_{\theta,t}(x_n) \quad (27)$$

where y_n denotes the label generated by the teacher model for sample x_n , ϵ denotes the label smoothing coefficient, and C is the number of classes. In our experiments, the label smoothing coefficient is set to 0.1.

Uncertainty Elicitation We adopt a self-probing approach (see references) to obtain the teacher model’s uncertainty. During knowledge distillation, we train the student model to explicitly represent its internal confidence via a multi-task learning framework. The overall loss function $\mathcal{L}_{\text{UE}}$ is computed as:

$$\mathcal{L}_{\text{UE}}(\theta) = \mathcal{L}_{\text{expl}}(\theta) + \mathcal{L}_{\text{label}}(\theta) + \mathcal{L}_{\text{uncer.}}(\theta) \quad (28)$$

Debate Prompt

You are in a debate now, this debate about evaluating the Correctness of Answer in relation to the question. My opinion is not always true, you can ignore any incorrect part of my opinion. And you can refer to my opinion to revise your choice or defend your own. Please remember there should and must be a more plausible answer in the choices, and Correctness is a value between 0 and 100.

Figure 3: Prompt Used in the Debate Stage.

where $\mathcal{L}_{expl}$, $\mathcal{L}_{label}$, and $\mathcal{L}_{uncer}$ denote the loss functions for explanation generation, answer prediction, and uncertainty estimation, respectively, all of which are formulated using the standard cross-entropy loss.

Consistency and Language Calibration based Filtering We remove samples whose teacher model’s self-consistency score or language calibration score is lower than 0.5. The student model is then trained on the filtered subset of samples.

Scenarios	Method	T5			LLaMA-1B		
		acc	ECE _{mass} ↓	ECE _{interval} ↓	acc	ECE _{mass} ↓	ECE _{interval} ↓
answer-only	5 CoTs	53.50	32.47	18.95	70.50	25.39	26.71
	LS	51.00	32.53	18.50	74.00	26.92	16.80
	UIKD _{SC}	53.00	30.21	16.32	73.50	30.21	18.49
	UIKD _{LC}	53.50	31.97	19.78	77.00	29.79	15.81
	UIKD	54.00	29.79	13.31	77.50	28.51	14.14
CoT	5 CoTs	57.00	35.94	20.25	73.50	19.60	9.71
	UIKD _{SC}	54.00	33.51	19.87	70.00	30.08	17.69
	UIKD _{LC}	56.50	34.33	24.81	67.50	28.79	15.81
	UIKD	58.50	33.41	19.21	71.50	26.95	14.11

Table 12: The experimental results on SVAMP.

F Experimental Results on SVAMP

We performed preliminary experiments on the SVAMP (Patel et al., 2021), as shown in Table 12. On the T5 model, UIKD improves both calibration and accuracy, reducing ECE_{mass} by 2.68 in the answer-only scenarios. However, on the LLaMA-1B, our method leads to an increase in the ECE. A possible reason is that the long generation process accumulates excessive uncertainty, resulting in insufficient confidence for the model.

G Details of the debate

Figure 3 presents the prompts used in our debate process.

H Few-shot Prompt

Table 13 presents the 5-shot prompts used in our process of generating data with teacher model GPT-3.5-turbo.

I Reliability Analysis

Figure 4 displays reliability diagrams for T5 models on the CQA dataset (answer-only setting). The plots confirm that UIKD effectively mitigates both overconfidence and underconfidence across the spectrum, yielding superior calibration compared to baselines.

J Case Study

Table 14 presents case studies from the CQA and ANLI datasets. In the first example, although both the baseline and UIKD make incorrect predictions, UIKD assigns a substantially lower confidence score, indicating reduced overconfidence. In the second and third examples, UIKD corrects the baseline errors and assigns high confidence to the correct answers. These examples show that UIKD produces confidence that better align with prediction correctness, lowering confidence for erroneous outputs while maintaining high confidence for correct predictions.

K High-Confidence Wrong-Consensus Cases

Since UIKD relies on debate-derived correctness scores to estimate the reliability of distillation samples, we further analyze a failure mode in which both debate agents assign high correctness scores to an incorrect answer. Table 15 presents representative high-confidence wrong-consensus examples, where both debate agents assign correctness scores of at least 80. In CQA, such errors mainly occur for incorrect options that are semantically close to the gold answer. In ANLI, such errors mainly arise when both debate agents rely on the same common-sense knowledge, but this knowledge conflicts with the given premise.

Task	Few-shot Prompt
CQA	Example 1: A beaver is known for building prowess, their supplies come from where? Answer Choices: (a) british columbia (b) body of water (c) wooded area (d) pay debts (e) zoo. Reason: Beavers are known for building dams and lodges using materials like branches, twigs, and mud, which are typically found in wooded areas. The answer is (c) wooded area. Example 2: A car was hailed to chauffeur someone to the opera house, where was it heading? Answer Choices: (a) go downtown (b) appear suddenly (c) go fast (d) bottom out (e) east. Reason: Opera houses are typically located in downtown areas. The answer is (a) go downtown. Example 3: A child wants to play, what would they likely want? Answer Choices: (a) fall down (b) breathe (c) play tag (d) be dismembered by a chainsaw (e) become adult. Reason: Children typically want to engage in games such as tag. The answer is (c) play tag. Example 4: A farmer sees a weasel in the woods, where is the farmer? Answer Choices: (a) chicken coop (b) beach (c) fairytale (d) great outdoors (e) corn fields. Reason: The presence of a weasel in the woods suggests a natural environment. The answer is (d) great outdoors. Example 5: A gentleman is carrying equipment for golf, what is he likely to have? Answer Choices: (a) club (b) assembly hall (c) meditation center (d) meeting (e) church. Reason: Golf equipment typically includes a golf club. The answer is (a) club.
ANLI	Example 1: Premise: The Parma trolleybus system comprises four urban routes. Hypothesis: The trolleybus system has over 2 urban routes. Reason: Four routes imply more than two. The answer is entailment. Example 2: Premise: Paul Beard was an English violinist and also a teacher. Hypothesis: Paul Beard was a teacher most of his life. Reason: The premise does not specify this. The answer is neutral. Example 3: Premise: Jesse James was among pro-Confederate guerrillas. Hypothesis: Jesse James was a guerrilla in the Union army. Reason: This contradicts the premise. The answer is contradiction. Example 4: Premise: Will Vodery was a Broadway composer and conductor. Hypothesis: Will Vodery conducted most Broadway shows. Reason: This claim is not supported. The answer is neutral. Example 5: Premise: Rhodochiton is native to southern Mexico and Guatemala. Hypothesis: Rhodochiton atrosanguineus is native to more than one country. Reason: The premise supports this. The answer is entailment.

Table 13: Few-shot prompting examples used for CQA and ANLI.

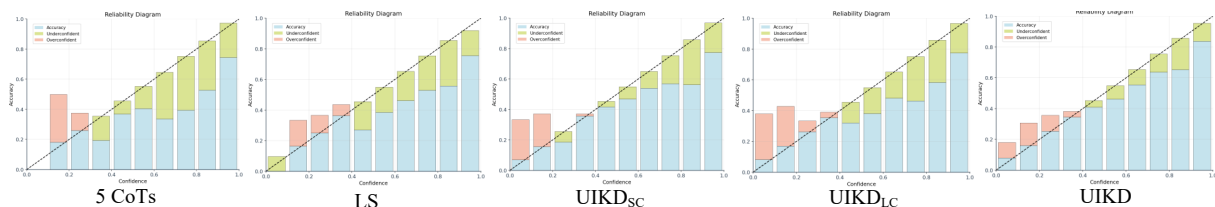


Figure 4: Reliability Diagrams of Student Models Obtained by Different Methods.

Dataset	Model Input	5 CoTs		LS		UIKD _{sc}		UIKD _{LC}		UIKD	
		answer	confidence	answer	confidence	answer	confidence	answer	confidence	answer	confidence
CommonsenseQA	Where would you put a glass after drinking from it? (a) ocean (b) water cooler (c) cabinet (d) dishwasher (e) dining room	cabinet	91.82%	cabinet	92.27%	cabinet	82.68%	cabinet	66.62%	cabinet	30.52%
	Billy set aside a block of time for having fun after work. Why might he do this? (a) happiness (b) stress relief (c) pleasure (d) ocean (e) may laugh	may laugh	72.71%	may laugh	89.14%	stress relief	28.18%	stress relief	73.90%	stress relief	91.45%
ANLI	Susan Sadlowski Garza is a member of the Chicago City Council serving as Alderman for the 10th ward. The 10th ward is located on Chicago's southeast side and includes East Side, Hegewisch, Jeffrey Manor, South Chicago and South Deering. She is serving her first term after defeating Rahm Emanuel ally John Pope in the 2015 election. </s> Susan Sadlowski Garza never met Rahm Emanuel.	contradiction	57.59%	contradiction	69.81%	neutral	53.37%	neutral	80.28%	neutral	92.21%

Table 14: Case studies comparing predicted answers and confidence scores across different distillation methods. Correct answers appear in red.

Dataset	Question	Gold bel	La- sensus	Wrong Con-	Reason
CQA	He put in a mighty effort to his solution, but a better one was discovered, making his what? Choices: unimportant, innovative, insignificant, helpless, powerless.	unimportant	insignificant		The phrase "a better one was discovered" suggests that the initial effort was not the best solution. This indicates that the initial effort was rendered "insignificant" or less important in comparison to the newly discovered solution.
ANLI	Premise: Texas District 6 of the United States House of Representatives is a Congressional district that serves an area including four counties to the south of the Dallas/Fort Worth area plus the southeast corner of Tarrant County. As of the 2000 census, District 6 represents 651,620 people. The current Representative from District 6 is Joe Barton and has been since 1985. Hypothesis: The current Representative from District 6 is Democrat Joe Barton and has been since 1985.	neutral		contradiction	Joe Barton is a member of the Republican Party, not the Democratic Party, which makes the information in the hypothesis about him being a Democrat incorrect. Therefore, the hypothesis directly contradicts the information provided in the premise. The answer is contradiction.

Table 15: Examples of high-confidence wrong-consensus cases in CQA and ANLI.