FIRST: Teach A Reliable Large Language Model Through Efficient Trustworthy Distillation

Kashun Shum[♡*], Minrui Xu[♡*], Jianshu Zhang[♠*], Zixin Chen[♡], Shizhe Diao [♦]
Hanze Dong[♡], Jipeng Zhang [♡], Muhammad Omer Raza [♣]
[♡]The Hong Kong University of Science and Technology, [♠]Wuhan University
[♦]NVIDIA, [♣] Purdue University
{ksshumab, mxubh}@connect.ust.hk

Abstract

Large language models (LLMs) have become increasingly prevalent in our daily lives, leading to an expectation for LLMs to be **trustworthy** — both accurate and well-calibrated (the prediction confidence should align with its ground truth correctness likelihood). Nowadays, fine-tuning has become the most popular method for adapting a model to practical usage by significantly increasing accuracy on downstream tasks. Despite the great accuracy it achieves, we found fine-tuning is still far away from satisfactory trustworthiness due to "tuning-induced mis-calibration". In this paper, we delve deeply into why and how mis-calibration exists in fine-tuned models, and how distillation can alleviate the issue. Then we further propose a brand new method named E**F**f**I**cient T**R**ustworthy Di**ST**illation (**FIRST**), which utilizes a small portion of teacher's knowledge to obtain a reliable language model in a cost-efficient way. Specifically, we identify the "concentrated knowledge" phenomenon during distillation, which can significantly reduce the computational burden. Then we apply a "trustworthy maximization" process to optimize the utilization of this small portion of concentrated knowledge before transferring it to the student. Experimental results demonstrate the effectiveness of our method, where better accuracy (+2.3%) and less mis-calibration (-10%) are achieved on average across both in-domain and out-of-domain scenarios, indicating better trustworthiness.[1]

1 Introduction

With the rapid development of large language models (LLMs), many powerful models have been deployed into our daily lives for practical usage to help us make decisions (Yao et al., 2023; Sha et al., 2023; Zhao et al., 2024). This makes it urgent for us to know to what extent we can trust the outputs of the models. Calibration is one of the most important indicators beyond accuracy, which provides a confidence measure to the model's predictions (Guo et al., 2017; Hsieh et al., 2023). In LLMs, confidence is exactly the probability for each generated token. Therefore, a well-calibrated model should align its prediction confidence with its ground-truth correctness likelihood as shown in Figure 1. As an example, recent hallucination detection methods rely on model prediction confidence as a significant indicator of potential hallucination (Zhang et al., 2023; Varshney et al., 2023). If the model is incapable of giving accurate confidence levels, people may fail to detect hallucinations due to the model's over-confidence, or people may falsely identify hallucinations due to the model's under-confidence. Mis-calibration brings significant challenges for the deployment of LLMs in real-world applications.

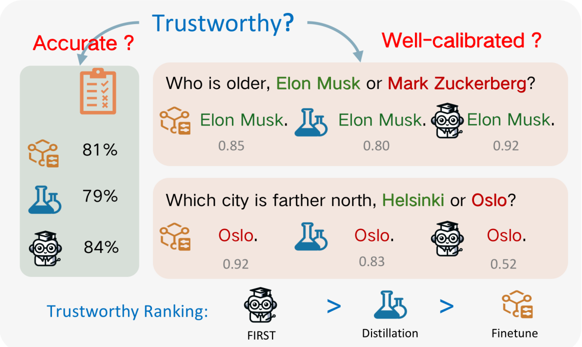


Figure 1: A trustworthy model should be both accurate (left) and well-calibrated (right). A well-calibrated model should produce high probabilities for the **correct answer** and low probabilities for the **wrong answer**.

Currently, there are two methods to obtain a language model for practical usage. First, fine-tuning, which fine-tunes pre-trained LLMs on specific datasets by matching each token entry with a target ground truth token. Although fine-tuning can consistently improve performance on downstream

[*] Equal Contribution.

[1] The code is available at https://github.com/SHUMKASHUN/FIRST.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 12646–12659
November 12-16, 2024 ©2024 Association for Computational Linguistics

tasks (Dodge et al., 2020; Sun et al., 2020; Ziegler et al., 2020), we identify that the model obtained in this way exhibits a nature of "tuning-induced mis-calibration". Second, distillation-based methods transfer knowledge (e.g., soft labels) from larger LLMs to smaller models (Gu et al., 2023). Although distillation shows better calibration than fine-tuning as it matches each token entry with a probability distribution instead of a hard label, we find it is still biased because of the mis-calibration nature of teacher models. In addition, distillation faces the challenge of determining the optimal amount of knowledge to transfer. Transferring all the teacher's knowledge leads to high computational costs while transferring too little knowledge results in poor accuracy. Therefore, it is crucial to balance between trustworthiness (accuracy and well-calibration) and efficiency for distillation-based methods.

To address the challenge of obtaining a trustworthy model, we propose e**F**f**I**cient t**R**ustworthy dis**T**illation (**FIRST**), aiming to efficiently utilize a relatively small amount of the teacher's knowledge. Specifically, we first identify the "concentrated knowledge" phenomenon, which shows that in the context of LLMs, the probability distribution of generated tokens is not uniform but rather concentrated on a few high-probability tokens. Based on this finding, we propose to use the top-5 tokens as the knowledge to balance the trade-off between storage space and the amount of knowledge transferred, achieving efficient distillation. Afterward, to eliminate the "tuning-induced mis-calibration" of the teacher model, we applied a "trustworthy maximization" to this portion of knowledge, ensuring that it maximizes the enhancement of the student model's accuracy while also guaranteeing its well-calibration.

We first validate our method in in-domain scenarios, discovering that the models obtained by FIRST achieve excellent accuracy, even with the use of a relatively small amount of top-5 knowledge and the "trustworthy maximization" process can significantly enhance these models' robustness to mis-calibration. Furthermore, we test our approach in out-of-domain settings, demonstrating that models obtained by FIRST still exhibit the best trustworthiness and hold generalization ability. This indicates that FIRST enables smaller models to genuinely learn the capability of being trustworthy, rather than being confined to in-domain scenarios.

In summary, our key contributions include:

(i) We discover that LLMs exhibit "concentrated knowledge" and "tuning-induced mis-calibration" phenomena, providing insights into obtaining trustworthy models.

(ii) We propose **FIRST**, which maximizes the effectiveness and trustworthiness of a relatively small portion of knowledge transferred from the teacher by "trustworthy maximization" to obtain a trustworthy student model.

(iii) Extensive experiments demonstrate that models obtained using FIRST consistently achieve the highest level of trustworthiness across different settings.

2 Related Work

2.1 Trustworthy Models

The current evaluation of LLMs predominantly focuses on accuracy, overlooking whether the models truly know the answer or are merely guessing (i.e. trustworthy). Recent works (Sun et al., 2024; Steyvers et al., 2024) have demonstrated that accurate LLMs may not necessarily be "trustworthy" due to a significant calibration gap, so-called mis-calibration. This gap prevents us from trusting the output of the models, and it can further cause LLMs to generate harmful content, especially when subjected to adversarial attacks or jailbreak prompts (Mo et al., 2024; Yao et al., 2024). Our work further reveals how mis-calibration exists in different tuning methods and proposes a new trustworthy evaluation metric that covers both accuracy and calibration.

To achieve a well-calibrated LLM, recent work shows soft-label distillation shows better calibration ability (Gu et al., 2023). However, it still suffers from biased labels due to the mis-calibration nature of the fine-tuned teacher model. Our work is an improvement on this line of work by applying "concentrated knowledge" and "trustworthy maximization", leading to better accuracy, efficiency, and trustworthy.

2.2 Knowledge Distillation

Knowledge Distillation is a form of transfer learning that facilitates the transfer of knowledge from a larger teacher model to a smaller student model. The goal is to reduce the model size while maintaining or even improving performance. Based on

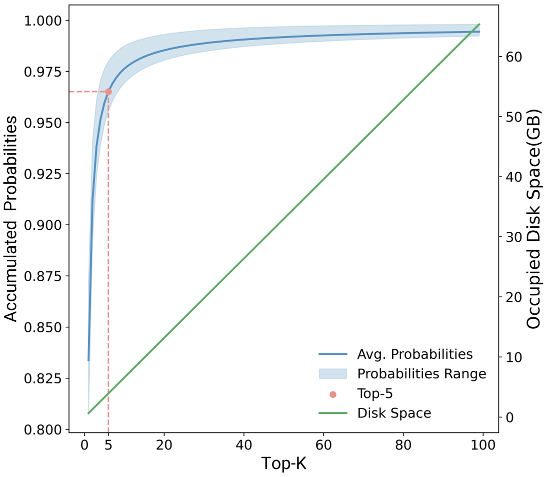


Figure 2: The blue line with range shows the averaged accumulated probability coverage for each token entry, from Top-1 to Top-100. **"Concentrated Knowledge"** : The red point represents accumulated probability for Top-5 tokens already exceed 95%. The green line describes the disk usage if use Top-K token distribution during distillation.

whether we can access prediction probability, the existing distillation methods can be categorized into two types: Black-box Distillation and White-box Distillation.

Black-box Distillation refers to distillation from models that we are unable to access the weight and prediction logits such as PaLM (Chowdhery et al., 2022). Recent studies have attempted to distill reasoning ability from GPT (Ho et al., 2023; Shridhar et al., 2023) or some emergent ability such as chain-of-thought (Hsieh et al., 2023; Li et al., 2023). However, these methods may still be categorized as the genre of data-augmentation-and-then-fine-tuning approaches.

White-box Distillation means the teacher models are either fully open-sourced such as Llama (Touvron et al., 2023a) or they can return partial probability distribution of the generated tokens, such as code-davinci-002. Instead of the hard token fine-tuning, white-box distillation typically uses more fine-grained signals by matching a distribution between teachers and students (Gu et al., 2023; Latif et al., 2023; Agarwal et al., 2024). Further, in the field of white-box distillation, there are two different ways: online distillation and offline distillation. Online distillation (Gu et al., 2023; Zhou et al., 2023) needs to keep both the teacher model and the student model on the GPU simultaneously during training. On the other hand, offline distillation typically involves obtaining knowledge from the teacher model beforehand. Our work is an extension of white-box offline distillation and focuses on how white-box offline distillation can be improved in terms of trustworthiness by re-calibrating the teacher distribution.

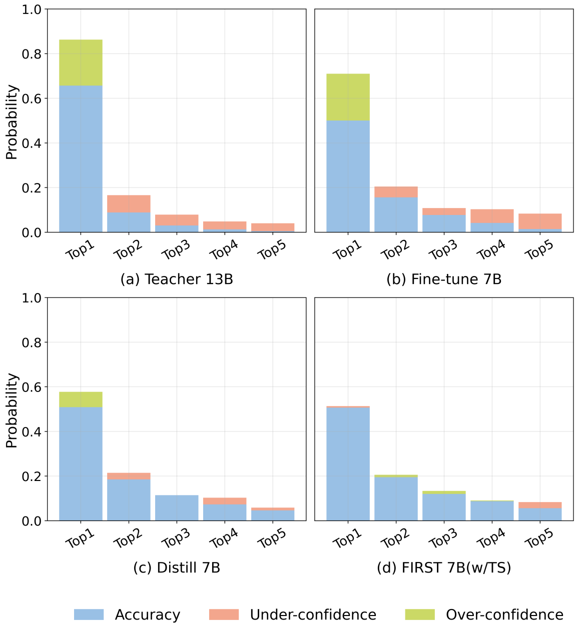


Figure 3: **"Tuning-induced Mis-calibration"** : Position-wise prediction probabilities with corresponding actual accuracy of (a) fine-tuned teacher model and (b) fine-tuned small model, (c) distilled model and (d) model produced by FIRST.

3 Preliminaries

3.1 Concentrated Knowledge

In the process of searching for a suitable trade-off between the amount of knowledge to transfer from the teacher model and efficiency, we begin by visualizing the probability distribution for each token entry. As illustrated in Figure 2, the blue line with range describes how averaged accumulated probabilities increase when we select more tokens (ranked from highest probability to lowest probability in one entry). The trend clearly shows a few top-position tokens take most of the probability information of a token entry. To be specific, the accumulated probabilities of top-5 tokens can occupy over 95% probabilities while the remaining 49995 (i.e. a model with vocab. size of 50k) tokens have nearly 0 probability. We named this phenomenon "Concentrated Knowledge" as almost full knowledge of a token entry is stored in its top-k tokens where the remaining tokens have negligible information.

3.2 Tuning-induced Mis-calibration

In the context of LLMs, mis-calibration can be divided into two types: over-confidence and under-confidence. Over-confidence occurs when the predicted probability of a token is higher than its actual accuracy, while under-confidence takes place when the predicted probability is lower than the actual accuracy.

During the fine-tuning process of LLMs, cross-entropy loss is commonly employed, which encourages the models to assign a probability of 1 to one token and 0 to all other tokens based on the ground-truth token. This training nature results in 1.) an over-estimation of the ground truth token's probability and 2.) an under-estimation of all other token's probability. As shown in Figure 3 (a) and (b), it is observed that both fine-tuned LLMs exhibit over-confidence in their top-1 token predictions, while demonstrating under-confidence in the subsequent tokens. This phenomenon, which we call "tuning-induced calibration", highlights the untrustworthy nature of fine-tuned models.

Since fine-tuned teacher models suffer from this tuning-induced mis-calibration, if the knowledge from the mis-calibrated teacher models is directly used in traditional distillation-based methods, the student models are very likely to inherit the same mis-calibration nature as depicted in Figure 3 (c). Motivated by the tuning-induced mis-calibration, our proposed method incorporates a "trustworthy maximization" procedure to re-calibrate the knowledge derived from the teacher models. This enables us to obtain a genuinely trustworthy student model.

3.3 Expected Calibration Error

To measure calibration in the context of LLMs, we adapt the expected calibration error (ECE) to the free-text generation task by treating the generation of a single token as a classification task. In this adaptation, we restrict the model to generate only one token from a set of candidate choices (e.g., A/B/C/D). For each token, we obtain the highest probability choice using $\arg\max_{i \in C} P(i)$, where C represents the set of candidates. The probability of the chosen token is taken as the predicted confidence, and we calculate the accuracy by comparing the predicted choice to the ground truth. Then we utilize a total M probability intervals as bins and categorize each chosen token into m-th bin according to the predicted confidence. The ECE (Guo et al., 2017) can be computed as follows:

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)| \quad (1)$$

Here, M is the number of bins. B_m represents the set of predictions in bin m, $|B_m|$ is the number of prediction instances in bin m, and n is the total number of predictions. $acc(B_m)$ is the average accuracy of predictions in bin m, and $conf(B_m)$ is the average confidence of predictions in bin m. A lower ECE value indicates that the model's predicted probabilities are more consistent with actual outcomes, meaning the model is better calibrated.

3.4 Trustworthy Score

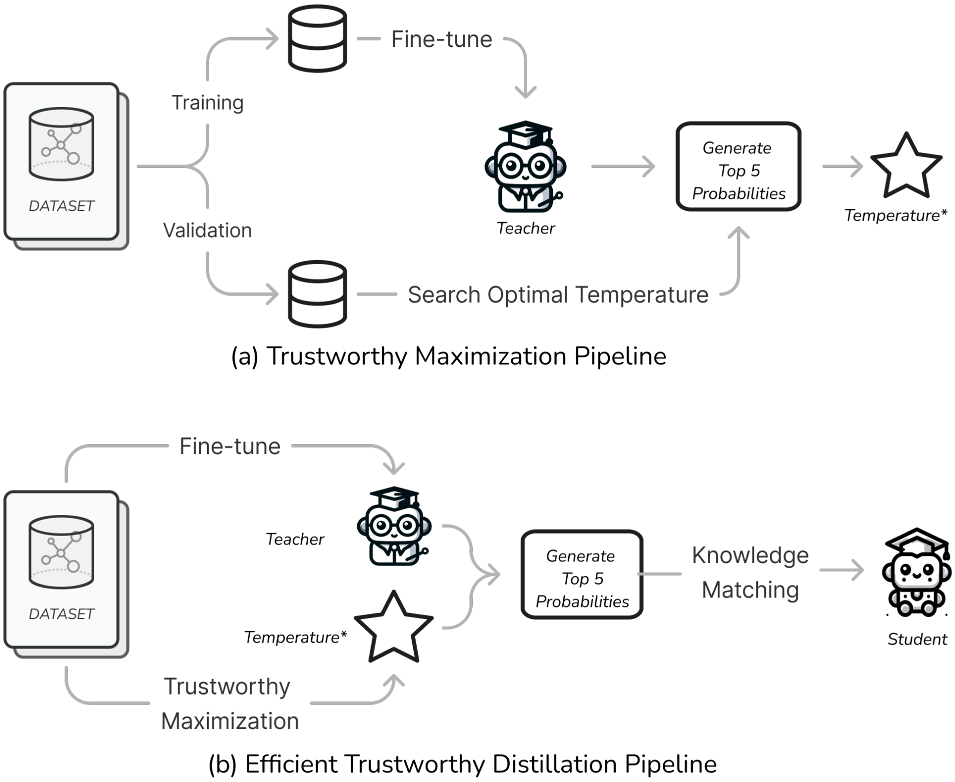
When evaluating the trustworthiness of a model, it is essential to consider both high accuracy and effective calibration. Existing benchmarks primarily focus on accuracy, assuming that higher accuracy implies greater trustworthiness. However, our discovery of the widespread issue of "tuning-induced mis-calibration" has highlighted the inadequacy of relying solely on accuracy for a comprehensive evaluation of model trustworthiness. To address this limitation, we propose Trust Score metric to quantify a model's trustworthiness, which considers two key aspects: its ability to provide accurate answers (measured by Acc) and its capacity to align predicted confidences with actual accuracies (measured by ECE). The Trust Score is defined as follows:

$$Trust = Acc - ECE \quad (2)$$

By incorporating the Trust Score, we achieve a more balanced evaluation of trustworthiness, taking into account both accuracy and calibration.

4 Efficient Trustworthy Distillation

In this section, we introduce e**F**f**I**cient t**R**ustworthy dis**T**illation (**FIRST**), which can be divided into three parts. Firstly, we select top-5 tokens as knowledge for transfer (Efficient Knowledge Selection) in Sec. §4.1. Then, we adjust the knowledge for trustworthiness to ensure that the subsequent smaller models can maximize its utility (Knowledge Trustworthy Maximization) in Sec. §4.2. Finally, we describe the learning process of the student model (Knowledge Matching) in Sec. §4.3. The overall pipeline is shown in Figrue 4.



(a) Trustworthy Maximization Pipeline

(b) Efficient Trustworthy Distillation Pipeline

Figure 4: (a) The Trustworthy Maximization Step: we first fine-tune our the teacher model and then generate top-5 probabilities of all tokens and run a grid search to select the optimal temperature based on the validation set. (b) The overall Efficient Trustworthy Distillation Pipeline: based on the selected optimal temperature from (a), we obtain a well-calibrated student model by knowledge matching between student's knowledge and the portion of teacher knowledge.

4.1 Efficient Knowledge Selection

Transferring knowledge directly from teachers to students can be computationally costly and storage-intensive. For example, if we consider a vocabulary size of 50,000 tokens, retrieving the complete probability distribution from a dataset of 100,000 samples, with an average length of 2,048, would require a staggering 120 TB of storage, which is impractical.

Based on the discovery of "concentrated knowledge" in teacher LLMs, we observe that the majority of knowledge is concentrated within a small portion of top-position tokens, as elaborated in Section §3.1. Therefore, considering that both computation and disk space increase linearly with the number of selected token entries, we argue that it is not necessary to use the complete probability distribution. Instead, by selecting a small amount of top-position tokens that contain majority of knowledge, we can strike the optimal balance between computational overhead and effectiveness. As depicted in Figure 2, accumulated probability of top-5 token entries occupy more than 95% probabilities while reducing storage from 120 TB to 1.2 GB.

4.2 Trustworthy Maximization

Once the top-5 tokens and their corresponding probabilities are collected from the teacher model, it is crucial to subject this knowledge to further processing to ensure proper calibration, as teacher models can also suffer from "tuning-induced mis-calibration" due to fine-tuning (as we elaborate in Sec. §3.2). This additional calibration step ensures that the student model improves in both accuracy and trustworthiness.

Label Smoothing: Similar to Müller et al. (2019), we first attempted to address tuning-induced mis-calibration" by applying a smoothing coefficient, denoted as δ, to mitigate the teacher model's over-confidence in its top-1 token predictions while alleviating under-confidence in other predicted tokens as follows:

$$\begin{cases} P_T(i) := P_T(i) - \delta & \text{if } i = 1 \\ P_T(i) := P_T(i) + \frac{\delta}{4} & \text{if } 2 \leq i \leq 5 \end{cases} \quad (3)$$

Here, T denotes the teacher model, $P_T(i)$ represents the probability of the i-th top token. While label smoothing can effectively mitigate over-confidence in top-1 token predictions, we have identified significant drawbacks associated with this approach. Firstly, directly applying label smoothing may compromise the preservation of token rankings, particularly between the top-1 and top-2 tokens. This can lead to a decline in model performance in certain cases. Secondly, label smoothing uses a constant probability, disregarding the varying levels of over-confidence or under-confidence in different token entries. Consequently, this can result in a transition from under-confidence to over-confidence among the top 2-5 tokens, making it challenging to achieve a balanced calibration across all of them.

Temperature Scaling: Subsequently, we explore another approach using a temperature scaling technique (Guo et al., 2017) to re-calibrate the probabilities:

$$P_T(i) = \frac{\exp(P_T(i)/c)}{\sum_j \exp(P_T(j)/c)} \tag{4}$$

This method offers several advantages. First, it allows for a more fine-grained adjustment of the probability distribution by controlling the temperature scaling parameter c, which can be optimized to achieve the lowest ECE values. Second, unlike label smoothing, temperature scaling can effectively balance the confidence levels of both top-1 and subsequent tokens, reducing both over-confidence and under-confidence issues by preserving token rankings and avoiding transition between under-confidence and over-confidence.

This results in a more consistent and reliable calibration across all tokens, thereby enhancing the overall trustworthiness of the knowledge. Additionally, we find that selecting the optimal c parameter on the validation set to maximize the knowledge can significantly enhance the effectiveness of transferring trustworthy knowledge. The knowledge processed by using this c yields the best results for the student model (detailed in Sec. §5.5). Due to the low cost of selecting c on the validation set, we can tailor different c values for different tasks. This demonstrates "temperature scaling" excellent scalability and flexibility.

4.3 Knowledge Matching

After obtaining the re-calibrated probability data P_T that contains $P_T(1), P_T(2), \ldots, P_T(5)$, we use the same training data to train the student model. Instead of utilizing language modeling loss on hard labels, the probabilities of the 5 tokens that correspond to the teacher's top-5 of the student model are retrieved as P_S which contains $P_S(1), P_S(2), ..., P_S(5)$. Kullback–Leibler divergence is then used to measure the loss between the teacher model and the student model:

$$Loss(y_{1:N}) = \sum_{t=1}^{N} D_{KL}(P_T||P_S) \tag{5}$$

5 Experiment

5.1 Experimental Settings

Our experiments focus on both In-Domain and Out-of-Domain settings to ensure generalization abilities. In the **In-Domain setting**, we utilize CommonsenseQA (CSQA) (Talmor et al., 2019) and BoolQ (Clark et al., 2019) for both training and testing. In the **Out-of-Domain setting**, we fine-tune and distill smaller models on a commonly used instruction-following dataset, Alpaca (Taori et al., 2023), while, testing the models' performance over unseen task CommonsenseQA (CSQA) and OpenBook QA (OBQA) (Mihaylov et al., 2018). This approach allows us to assess the generalization abilities of the smaller models on unseen tasks, simulating real-world scenarios where these models need to perform on unfamiliar tasks.

To ensure the practicality of our approach, we select three widely used model families for our experiments: Llama-1 (Touvron et al., 2023a), Llama-2 (Touvron et al., 2023b), and OpenLlama (Geng and Liu, 2023). In our experiments, we test four types of smaller models obtained through different methods:

1) **Fine-tune** $_{\textbf{7B}}$: Obtained by using fine-tuning with hard labels.

2) **Distill** $_{\textbf{7B}}$: Obtained by distillation methods without "knowledge trustworthy maximization". For a fair comparison with our approach, we also use the top-5 tokens as knowledge in the latter comparison.

3) **FIRST** $_{\textbf{7B w/TS}}$: Obtained by our proposed method, primarily using temperature scaling (TS, see Eq. 4) within the trustworthy maximization phase.

4) **Distill** $_{\textbf{7B w/ LS}}$: We also explore the use of label smoothing (LS, see Eq. 3) to show why we ultimately adopt TS over LS in "knowledge trustworthy maximization". In the latter experiments,

	IN-DOMAIN						OUT-OF-DOMAIN					
	CSQA			BoolQ			CSQA			OBQA		
	$ECE \downarrow$	$Acc \uparrow$	$Trust \uparrow$	$ECE \downarrow$	$Acc \uparrow$	$Trust \uparrow$	$ECE \downarrow$	$Acc \uparrow$	$Trust \uparrow$	$ECE \downarrow$	$Acc \uparrow$	$Trust \uparrow$
	LLAMA 1 : 33B → 7B											
Teacher$_{\text{33B}}$	10.2	82.4	72.2	7.7	89.7	82	18.6	69.2	50.6	20.2	64.4	44.2
Fine-tune$_{\text{7B}}$	11.8	79.9	68.1	6.5	82.5	76	12.5	48.2	35.7	21.9	43.4	21.5
Distill$_{\text{7B}}$	9.4	78.9	69.5	**4.0**	85.3	81.3	5.3	43.1	37.8	18.1	39.8	21.7
Distill$_{\text{7B w/ LS}}$	9.1	78.1	69	19.0	85.3	66.3	5.2	43.9	38.7	19.0	37.6	18.6
FIRST$_{\text{7B w/ TS}}$	**2.9**	**80.8**	**77.9**	**4.0**	**85.7**	**81.7**	**4.6**	**50.0**	**45.4**	**7.1**	**47.2**	**40.1**
FIRST to Fine-tune	↑8.9	↑0.9	↑9.8	↑2.5	↑3.2	↑5.7	↑7.9	↑1.8	↑8.7	↑14.8	↑3.8	↑18.6
	LLAMA 2 : 13B → 7B											
Teacher$_{\text{13B}}$	12.0	81.6	69.6	6.8	89.7	82.9	20.8	65.7	44.9	28.7	58.3	29.9
Fine-tune$_{\text{7B}}$	14.0	76.8	62.8	8.4	87.5	79.1	21.2	50.0	28.8	30.1	45.6	15.5
Distill$_{\text{7B}}$	10.9	80.0	69.1	4.0	85.3	81.3	7.7	50.9	43.2	12.5	46.6	34.1
Distill$_{\text{7B w/ LS}}$	10.3	**80.4**	70.1	3.9	87.5	83.6	7.5	51.1	43.6	16.2	47.6	31.4
FIRST$_{\text{7B w/ TS}}$	**6.3**	80.3	**74**	**1.4**	**87.9**	**86.5**	**5.5**	**51.4**	**45.9**	**8.1**	**49.5**	**41.4**
FIRST to Fine-tune	↑7.7	↑3.5	↑11.2	↑7	↑0.4	↑7.4	↑15.7	↑1.4	↑17.1	↑22	↑3.9	↑25.9
	OPENLLAMA : 13B → 7B											
Teacher$_{\text{13B}}$	13.2	78.5	65.3	7.5	87.6	80.1	16.7	49.5	32.8	13.4	50.0	36.6
Fine-tune$_{\text{7B}}$	10.5	75.0	64.5	3.6	81.5	77.9	21.6	28.3	6.7	16.1	30.4	14.3
Distill$_{\text{7B}}$	9.2	75.2	66	6.2	83.8	77.6	9.7	27.7	18	13.7	29.8	16.1
Distill$_{\text{7B w/ LS}}$	9.6	74.5	65.9	3.3	83.3	80	4.1	29.2	25.1	14.2	29.8	15.6
FIRST$_{\text{7B w/ TS}}$	**5.0**	**77.2**	**72.2**	**2.7**	**84.7**	**82**	**2.9**	**30.5**	**27.6**	**8.2**	**30.8**	**22.6**
FIRST to Fine-tune	↑5.5	↑2.2	↑7.7	↑0.9	↑3.2	↑4.1	↑18.7	↑2.2	↑20.9	↑7.9	↑0.4	↑8.3

Table 1: Smaller models obtained by our method FIRST consistently achieves high accuracy Acc across various scenarios while maintaining a low expected calibration error ECE (see Eq. 1). The higher trust scores $Trust$ (see Eq. 2), the more trustworthy models are. Note that in the out-of-domain setting, we only obtain smaller models by fine-tuning or distilling on Alpaca, with CSQA and OBQA being unseen in this context, validating the generalizability of our approach. ↑ represents the larger the better while the ↓ means the smaller the better. **Bold** represents the best.

we pick up the popular smoothing coefficient 0.1 follow previous works (Müller et al., 2020).

Additionally, we also provide the performance of **Teacher** models. For further implementation details, please refer to the Appendix A.

5.2 Experiment Results

Based on the results shown in Table 1, we draw the following conclusions:

• **Fine-tuning lead to catastrophic mis-calibration**: We observed that although fine-tuned smaller models achieve relatively high accuracy in both in-domain and out-of-domain settings, their ECE values are notably high, resulting in overall low trust scores and lower reliability. This mis-calibration phenomenon is particularly pronounced in out-of-domain scenarios. For instance, we observe that the ECE of the model fine-tuned on OpenLllama 7B in the out-of-domain CSQA task reaches 21.6%, while its accuracy is only 28.3%, indicating that smaller models obtained through fine-tuning tend to be unreliable on tasks they have not been trained on. In real-world scenarios, when smaller models are privately deployed, they will inevitably encounter tasks they have not been trained for. In such cases, there would be a mismatch between their confidence and true likelihood. They might confidently provide incorrect answers and even continuously emphasize their incorrect responses, thereby misleading users. This clearly does not meet the criteria of a trustworthy model.

• **Distillation brings bad calibration as well**: Furthermore, distilled models without "Knowledge Trustworthy Maximization" show relatively bad calibration ability. For in-domain tasks, the distilled Llama-1 7B and Llama-2 7B have ECE values of 9.4% and 10.9% on CSQA, a mis-calibration level similar to fine-tuned models. And distilled model of OpenLlama shows even worse calibration than fine-tuned models on BoolQ. While for accuracy, it generally has an improvement over standard fine-tuning, but on some settings such as Llama-1 on CSQA, it also shows worse performance than fine-tuning. This suggests that direct distillation without further process the knowledge does not consistently lead to better calibration and performance.

• **Temperature Scaling outperforms Label Smoothing**: Here, we compare the results of different methods used in the "Knowledge Trustworthy Maximization" phase. It is evident that FIRST$_{\text{7B w/ TS}}$ performs significantly better than

Distill$_{7B\text{ w/ LS}}$. In the in-domain setting of BoolQ, the ECE values of FIRST$_{7B\text{ w/ LS}}$ astonishingly reached 19.0%, significantly worse than Distill$_{7B}$, which does not apply any additional processing to the knowledge. This highlights that LS cannot deliver stable performance across all scenarios. In contrast, FIRST$_{7B\text{ w/ TS}}$ consistently achieves lower ECE in both in-domain and out-of-domain scenarios. Additionally, they attain better accuracy in most cases, resulting in the highest Trust Scores.

5.3 Reliability Analysis

Reliability Diagrams. To enhance our analysis and facilitate better comparisons, we employ reliability diagrams in addition to metric-based evaluations. As depicted in Figure 5, the reliability diagrams are divided into 10 bins based on the model's confidence. The bars represent the expected accuracy within each bin, and the colors indicate whether the model is under-confident (red) or over-confident (green) within each bin. A perfectly calibrated model would have a straight diagonal line from the bottom left to the top right of such a diagram, indicating that the confidence level is exactly consistent with expected accuracy.

The Fine-tune$_{7B}$ model exhibits catastrophic mis-calibration, primarily characterized by over-confidence in its predictions. This means that the model tends to assign higher confidence levels to its predictions than what is justified by their actual accuracy. Although the Teacher$_{33B}$ model also suffers from over-confidence, its overall high accuracy results in a much higher trust score. Additionally, the Distill$_{7B}$ model demonstrates slightly improved calibration compared to the Fine-tune$_{7B}$ model. Remarkably, our FIRST$_{7B}$ model outperforms the other models, including the teacher model. It exhibits noticeably less under-confidence and over-confidence, as indicated by the smaller areas of the red and green bars, respectively, and its proximity to the perfect calibration line.

5.4 Analysis of Top-5 Selection.

Figure 2 illustrates the disk space usage and cumulative probability coverage for knowledge selection ranging from the top-1 to the top-100 tokens. The blue line represents the average accumulated probabilities, while the shaded area indicates the range of probabilities. The green line shows the corresponding disk space required. The reasons we finally adopted top-5 are as follows:

1. **Efficient Probability Coverage**: The figure demonstrates that selecting the top-5 tokens covers over 95% of the total probability. This high coverage ensures that the majority of relevant knowledge is captured, making the distillation process effective.

2. **Minimal Disk Space Usage**: The green line indicates the disk space required for storing the selected tokens. By selecting only the top-5 tokens, we significantly reduce the storage requirements compared to selecting more tokens. This efficiency is crucial for offline distillation, where disk space can be a limiting factor.

3. **Balancing Trade-offs**: The top-5 selection strikes a balance between maximizing probability coverage and minimizing disk space usage. This balance ensures that the distilled knowledge is both comprehensive and storage-efficient, enabling practical implementation in various scenarios.

4. **Scalability**: Our method exhibits strong scalability. It is naturally extendable to distillation from models such as the GPT-3 series (text-davinci-003), which can only return top-5 token probabilities. This increases the range of LLMs that can be used as teacher models, allowing student models to be effectively trained even in semi-black box scenarios.

5.5 Temperature Scaling Parameter Analysis

As described in the section on Knowledge Trustworthy Maximization (Sec. §4.2), we employ a temperature scaling parameter to optimize the ECE (Expected Calibration Error) value on the validation set, as illustrated in the left part of Figure 6. By employing grid search, we initially partition the range from 0 to 1 into increments of 0.1 and identify the temperature associated with the lowest ECE value, for instance, 0.3. A larger temperature results in all top-5 tokens converging to the same probabilities, specifically 0.2 when the number of candidate choices is 5. When the temperature is set to 1, the probability of the top-1 token is dramatically compressed, while the probabilities of the other tokens are enlarged accordingly. Conversely, a temperature of 0.1 can even amplify the probabilities of over-confident tokens, leading to even worse calibration.

To further refine the search for the optimal tem-

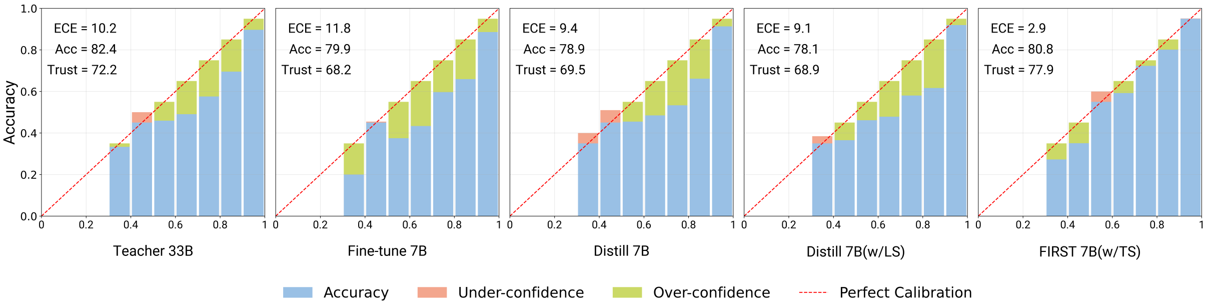


Figure 5: Reliability diagrams based on Llama-1 reveal the mis-calibration of various models on the CSQA dataset. In these diagrams, the X-axis is confidence divided into 10 bins, representing the model's confidence levels for each question's answer tokens. The Y-axis represents the accuracy within each bin. The red bar represents the degree to which the actual accuracy is higher than perfect calibration (under-confident), while the green bar means that the actual accuracy is lower than perfect calibration (over-confident).

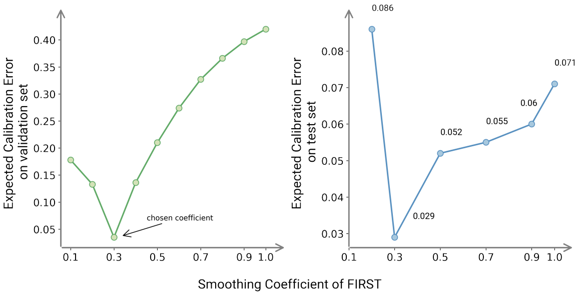


Figure 6: Left shows the comparison of different smoothing coefficients on the validation set, while the right part demonstrates its corresponding calibration effect on the test set.

perature, we narrow down the interval and use a smaller step size of 0.02. This allows us to pinpoint the best temperature more precisely. Additionally, we compare the performance of FIRST using the selected optimal temperature with other different temperatures as shown in the right part of Figure 6. FIRST with optimal temperature do outperform those with other levels of temperatures with a large margin, indicating the effectiveness of selecting such optimal temperature.

6 Conclusion

In conclusion, our proposed method, e**F**f**I**cient t**R**ustworthy di**ST**illation (FIRST), effectively enhances both accuracy and calibration in large language models. By applying "trustworthy maximization", FIRST efficiently transfers the minimal yet most effective knowledge from teacher to student models. Experimental results show that FIRST consistently improves trustworthiness across various scenarios, demonstrating its potential to create reliable language models for practical applications.

7 Limitations

It is shown that our efficient trustworthy distillation (FIRST) demonstrates superior calibration ability and performance over direct distillation and standard fine-tuning methods. However, despite these exciting results, there are still some limitations in our current work, as well as potential opportunities for future research.

Extend to Large Teacher Model : Due to the resource limitation, our largest teacher model is Llama 33B, which is not very large but already achieving exciting results by distillation to a 7B student model. We expect that employing a large teacher model such as 70B can lead to better calibration ability and performance since a larger model learns a better distribution. However, we are unable to explore how very large teachers perform due to resource limitations.

Top-K Chosen in Offline Distillation: Another limitation of this work is that it does not provide a rigorous study on how many token probabilities to choose for one entry is optimal for knowledge distillation in large language models. Currently, we consistently choose the top-5 token probability to retrieve because of the reasons stated in §5.4. However, how much token probability to use is optimal could be an important area for further exploration and development.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. 2024. On-policy distillation of language models: Learning from self-generated mistakes.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina. Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *NAACL*.

Shizhe Diao, Rui Pan, Hanze Dong, Ka Shun Shum, Jipeng Zhang, Wei Xiong, and Tong Zhang. 2023. Lmflow: An extensible toolkit for finetuning and inference of large foundation models.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping.

Xinyang Geng and Hao Liu. 2023. Openllama: An open reproduction of llama.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2023. Knowledge distillation of large language models.

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 - Volume 70*, ICML'17, page 1321–1330. JMLR.org.

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.

Shengding Hu, Yifan Luo, Huadong Wang, Xingyi Cheng, Zhiyuan Liu, and Maosong Sun. 2023. Won't get fooled again: Answering questions with false premises.

Ehsan Latif, Luyang Fang, Ping Ma, and Xiaoming Zhai. 2023. Knowledge distillation of llm for education. *arXiv preprint arXiv:2312.15842*.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023. Symbolic chain-of-thought distillation: Small models can also "think" step-by-step. *ArXiv*, abs/2306.14050.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Lingbo Mo, Boshi Wang, Muhao Chen, and Huan Sun. 2024. How trustworthy are open-source llms? an assessment under malicious demonstrations shows their vulnerabilities.

Rafael Müller, Simon Kornblith, and Geoffrey Hinton. 2019. *When Does Label Smoothing Help?* Curran Associates Inc., Red Hook, NY, USA.

Rafael Müller, Simon Kornblith, and Geoffrey Hinton. 2020. When does label smoothing help?

Hao Sha, Yao Mu, Yuxuan Jiang, Li Chen, Chenfeng Xu, Ping Luo, Shengbo Eben Li, Masayoshi Tomizuka, Wei Zhan, and Mingyu Ding. 2023. Languagempc: Large language models as decision makers for autonomous driving. *arXiv preprint arXiv:2310.03026*.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2023. Distilling reasoning capabilities into smaller language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7059–7073, Toronto, Canada. Association for Computational Linguistics.

KaShun Shum, Shizhe Diao, and Tong Zhang. 2023. Automatic prompt augmentation and selection with chain-of-thought from labeled data.

Mark Steyvers, Heliodoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu, Lukas Mayer, and Padhraic Smyth. 2024. The calibration gap between model and human confidence in large language models. *arXiv preprint arXiv:2401.13835*.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2020. How to fine-tune bert for text classification?

Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, et al. 2024. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *ArXiv*, abs/2307.03987.

Dongyu Yao, Jianshu Zhang, Ian G Harris, and Marcel Carlsson. 2024. Fuzzllm: A novel and universal fuzzing framework for proactively discovering jailbreak vulnerabilities in large language models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4485–4489. IEEE.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. 2023. Enhancing uncertainty-based hallucination detection with stronger focus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 915–932, Singapore. Association for Computational Linguistics.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2024. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19632–19642.

Yongchao Zhou, Kaifeng Lyu, Ankit Singh Rawat, Aditya Krishna Menon, Afshin Rostamizadeh, Sanjiv Kumar, Jean-François Kagy, and Rishabh Agarwal. 2023. Distillspec: Improving speculative decoding via knowledge distillation.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-tuning language models from human preferences.

	Standard Fine-tuning	Direct Distillation	FIRST
Question	Which city is farther north, Oslo or Helsinki?		
Correct Answer	Helsinki		
Generated Confidence	Oslo is farther north than Helsinki. 0.92 → over-confident	Oslo is farther north than Helsinki. 0.83 → over-confident	Oslo is farther north than Helsinki. 0.52
Question	Is Donald Trump a Neo-con American politician and businessman for the Republicans, with a long and varied career?		
Correct Answer	No		
Generated Confidence	Yes. 0.91 → over-confident	Yes. 0.85 → over-confident	Yes. 0.54
Question	If I want to visit Beijing in spring, when should I go? Answer Choices: (a) June (b) July (c) August (d) September (e) October		
Correct Answer	None		
Generated Confidence	(c). 0.58 → over-confident	(d). 0.41 → over-confident	(d). 0.27

Table 2: A case study on how fine-tuned model and direct distilled model tend to over-confident on the wrong answer with high confidence. While FIRST though outputs a wrong answer, it produces low confidence to show its uncertainty.

A Detailed Experimental Setting

A.1 Implementation Details

We train our models on 8 GPU (RTX A6000 48G) using the Adam optimizer with beta set to be [0.9, 0.999] and epsilon fixed to be 1e-6 and cosine annealing scheduler with a warm-up ratio of 0.03. For fine-tuning, we utilize LMFlow (Diao et al., 2023) package to obtain a well fine-tuned model by a standard 3-epoch training and control the batch size to be 32 on each GPU and the learning rate for teacher models to be 2e-5. Finally, for distillation, the batch size is set to 32 on each GPU and we train our model for 3 epochs, the last checkpoint is used for evaluation since it has the best performance.

In addition, when implementing distillation without re-calibration, we use the following normalization function to normalize the top 5 distribution and prevent the probability to be 0.

$$P_T(i) = \frac{P_T(j) + \delta}{\sum_j (P_T(j) + \delta)}$$

In our setting, i, j = 1, . . . , 5, representing the top-5 token probability and δ is a small shift amount that prevent the probability to be 0 after normalization. The δ is set to be 1e-6 to minimize the influence.

A.2 Prompt and Data Format

For question-answering tasks, we follow Shum et al. (2023)'s format and fine-tune the model in a zero-shot setting. For out-of-domain tasks, we directly follow Alpaca's (Taori et al., 2023) setting to obtain the fine-tuned model. The full prompt formats are shown in Table 3.

B Additional Analysis

B.1 Case Study

We further conduct three case studies to show that FIRST indeed helps mitigate mis-calibration in real-world question answering.

As shown in Table 2, we ask the models of three different tuning methods on Alpaca to answer the question: `which city is farther north, Oslo or Helsinki?` The correct answer is `Helsinki` and the wrong answer is `Oslo`. From the output confidence, we can see that standard fine-tuned models and direct distillation give high confidence in the wrong answer, which is far from satisfactory for trustworthy in real-world settings, especially when additional post-processing procedures were expected to be applied to filter wrong answers by identifying unconfident responses. In comparison, FIRST greatly mitigates this

CSQA

Q: The sanctions against the school were a punishing blow, and they seemed to what the efforts the school had made to change?
Answer Choices:
(a) ignore
(b) enforce
(c) authoritarian
(d) yell at
(e) avoid
A: The answer is (a).

OBQA

Q: food is a source of energy for what?
Answer Choices:
(A) waterfalls
(B) fires
(C) grass snakes
(D) mountains
A: The answer is (C).

Alpaca

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
For the given list of items, classify them into two categories.

Input:
Carrot, Apple, Pumpkin, Orange

Response:
Fruits: Apple, Orange
Vegetables: Carrot, Pumpkin

BoolQ

Below is an instruction that describes a task. Write a response that appropriately completes the request.

Instruction:
Read the input passage and answer the question: is windows movie maker part of windows essentials? Your answer should be Yes or No.

Input:
Windows Movie Maker (formerly known as Windows Live Movie Maker in Windows 7) is a discontinued video editing software by Microsoft. It is a part of Windows Essentials software suite and offers the ability to create and edit videos as well as to publish them on OneDrive, Facebook, Vimeo, YouTube, and Flickr.

Response:
Yes

Table 3: Examples of our prompts and data formats for four different datasets. The same formats are used across all models and experiments.

mis-calibration by producing a confidence of around 50% which indicates the model is not sure about the generated answer, allowing systems to filter those undesirable answers by a hard confidence threshold.

In the third case, we follow the FalseQA (Hu et al., 2023). In this case, all of the answer choices are expected to be wrong and models should output a confidence of 25% in the top-1 token to achieve minimal ECE value. That's why our FIRST shows best calibration in this case.

C Trust Score Design

Expected calibration error (ECE) is calculated by weighted average of difference between confidence and accuracy, which means accuracy and ECE are naturally in the same scale. Given that higher accuracy while lower ECE is better, it is intuitive and reasonable to define the trustworthy score by subtracting ECE from the accuracy. Besides, the product of ACC and ECE (e.g. $Trust\ Score = ACC \cdot (1 - ECE)$)

will introduce a factor of ACC to the ECE score : $ACC - ACC \cdot ECE.$ This would bring unfairness when comparing large models (high accuracy) with small models (low accuracy) because we expect ECE has the same importance when evaluating either high-accuracy model or low-accuracy model.