Calibrated Offline Knowledge Distillation for Large Language Model Training

Anonymous ACL submission

Abstract

Knowledge Distillation (KD) in large language models (LLMs) which involves training a small model to mimic the behaviour of a large model by matching their output distribution, has shown remarkable improvement in performance and efficiency over standard fine-tuning. Despite the great success of these methods, distilled student models are still suffering from catastrophic mis-calibration due to the over-confident nature of the teacher model. In this paper, we present a comprehensive study on the importance and necessity of re-calibration during soft-label-based distillation. We further propose a soft-label-based **C**alibrated **O**ffline knowledge **D**istillation **(COD)** pipeline that can effectively determine to what extent different token probability should be reduced or raised, resulting in a consistent distillation of a reliable model. Specifically, we start by re-calibrating the token probability distribution generated by the teacher model, by reducing the probability of over-confident tokens and raising the under-confident ones. Then we train a student model to fit the calibrated distribution. We conduct extensive experiments on both in-domain and out-of-domain settings by comparing calibrated distillation with non-calibrated distillation and standard fin-tuning over three popular open-sourced language model family (Llama-1, Llama-2, and OpenLlama). Experimental results demonstrate that re-calibration before distillation can greatly improve the reliability of the model (by 4.3% expected calibration error on average) and generally further boost the downstream performance (by 2.5% accuracy on average).

1 Introduction

With the rapid development of large language models (LLMs), the number of powerful pre-trained models has been skyrocketing, and the paradigm of pretrain-then-finetune has become a common method for people to adapt pre-trained models for downstream tasks (Wei et al., 2022). On top of that, recent studies show that knowledge distillation from fine-tuned large models can potentially achieve better performance than standard fine-tuning (Gu et al., 2023; Agarwal et al., 2024), rendering distillation a promising alternative in the scenarios of small model training. Nevertheless, despite the consistently improved performance on downstream tasks, these methods can still bring catastrophic mis-calibration problems (OpenAI, 2023) due to the over-confident nature of fine-tuned 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). In LLMs, confidence is exactly the probability for each generated token. As LLMs have been widely adopted in our daily lives now, it is crucial to understand the extent to which we can trust the answers they generate. In other words, the probability corresponding to the predicted token should reflect its ground truth correctness likelihood. 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 hallucina-

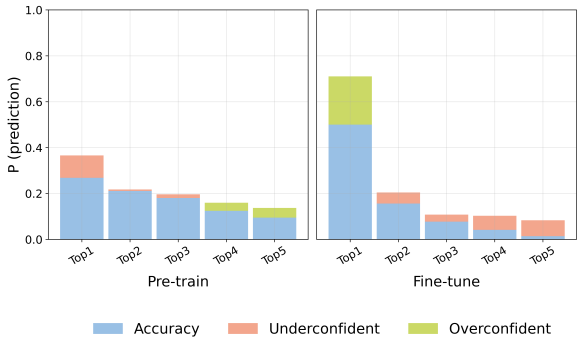


Figure 1: The position-wise confidence with its actual accuracy of pre-trained model and fine-tuned model. Fine-tuned models always be much over-confident on Top-1 token.

tions due to the model's over-confidence, or people may falsely identify hallucinations due to the model's under-confidence. This brings significant challenges for the deployment of LLMs in real-world applications.

In the process of searching for better ways to alleviate mis-calibration during distillation, we discover that both fine-tuned large and small models tend to be over-confident on the top-1 token and under-confident on top 2-5 tokens by employing a position-wise comparison on model predicted confidence and its actual accuracy as shown in Figure 1. The downside is clear from the depiction, as distillation by mimicking the not-well-calibrated distribution will result in a student model with great mis-calibration. In that case, re-calibration before distillation provides a promising way to adjust the teacher output probability distribution while preserving the ability of larger teacher models.

Motivated by this phenomenon, in this paper, we delve deeply into how re-calibration can affect the calibration and performance of distillation and first propose an efficient soft-label-based calibrated offline knowledge distillation pipeline for large language models named COD. Different from much previous hard distillation methods which utilize data generated from ChatGPT and then fine-tune student models on the generated data, our method utilizes logits of the teacher model and optimizes the student model using distribution match. Our pipeline mainly contains four steps as shown in Figure 2:

(1) **Teacher Building:** We first use domain data to supervised fine-tune a relatively large teacher model which is white-box to us.

(2) **Efficient Data Generation:** After obtaining the teacher model, we then let the teacher model generate the probability distribution for each token of the training dataset, and only keep the top-5 tokens for each token entry. This not only saves much disk space but also makes our pipeline fully compatible with GPT-3.5 series (text-davinci-003) which can only return top-5 token probabilities.

(3) **Re-calibration:** By pre-selecting a smoothing coefficient on validation set that can achieve best expected calibration error (ECE) score, we re-calibrate and normalize the Top-5 token probability of the teacher model in an offline manner.

(4) **Distribution Matching:** Finally, the soft labeled data are collected to teach student models. To be specific, we retrieve the 5 tokens of student model that corresponding to the top-5 tokens of the teacher model and optimize the student models by minimizing the Kullback-Leibler Divergence (KLD) (Kullback and Leibler, 1951) between the two probability distribution.

We conduct extensive experiments to compare distillation with and without re-calibration by employing several common baselines, demonstrating that catastrophic mis-calibration exists in both fine-tuning and distillation w/o re-calibration methods. On top of that, we explore how to pre-define a good smoothing coefficient which helps to determine to what extent different token probability should be reduced or raised in our COD method in order to distill a well-calibrated student model. Our experiments are based on Llama 1, Llama 2, and Open-Llama family (Touvron et al., 2023; Geng and Liu, 2023) since they are considered as most advanced open-sourced models so far and have a flexible range of model sizes. The results show that re-calibration can consistently improve the reliability of the distilled student model as well as improve the performance on downstream tasks in both in-domain and out-of-domain settings. Compared to direct distillation without re-calibration, our COD can generally improve 4.3% on ece, along with an averaged 2.5% increase in accuracy, showing the effectiveness of our proposed pipeline. In summary, our key contributions include:

(i) We show the surprising effectiveness and necessity of re-calibration in improving robustness to mis-calibration when compared with direct distillation and standard fine-tuning.

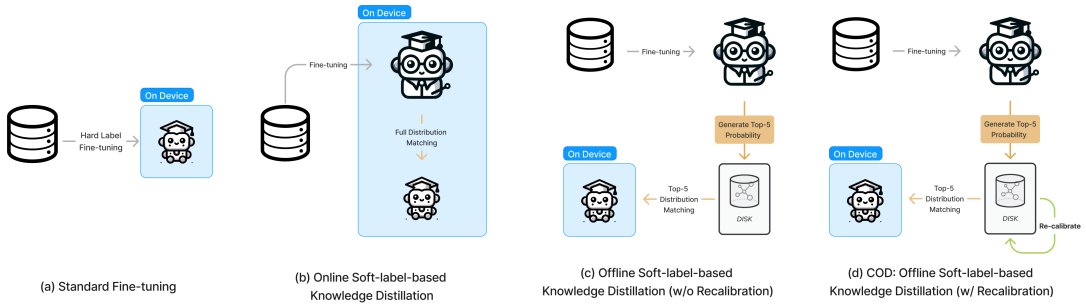
(ii) We proposed an efficient soft-label-based calibrated offline knowledge distillation (COD) pipeline for large language models which is scalable by allowing very large teacher models (e.g. >30B) and student models to be trained separated during distillation and can distill more reliable and stronger student model.

(iii) We conducted extensive experiments to quantify and analyze the benefits re-calibration brings in order to mitigate mis-calibration.

2 Related Work

2.1 Model Calibration

Calibration is a crucial aspect of modern neural network models, as it deals with predicting probability estimates that represent the true likelihood of



(a) Standard Fine-tuning

(b) Online Soft-label-based Knowledge Distillation

(c) Offline Soft-label-based Knowledge Distillation (w/o Recalibration)

(d) COD: Offline Soft-label-based Knowledge Distillation (w/ Recalibration)

Figure 2: The illustration and comparison of different student tuning pipelines. On Device indicates what model is placed on GPUs during training. (a) represents a standard fine-tuning pipeline which utilize hard label for training. (b)(c)(d) represent different distillation pipeline where (b) is under online setting which means teacher and student are both placed on GPU during distillation. Instead, our proposed COD in (d) make it offline by storing the teacher Top-5 distribution and further improve the pipeline by adding a re-calibration component.

correctness. It ensures that the confidence scores produced by models align with their actual performance. Recent studies have shown that modern neural networks are poorly calibrated (Guo et al., 2017), especially after fine-tuning (OpenAI, 2023) in large language model settings. To mitigate the issue, previous works contain unlearnable methods that heuristically manipulate the original confidence in predictions (Müller et al., 2019), and learnable methods that rely on extra calibration tasks which require extra data and more training cost (Chen et al., 2023). Different from previous literature, this work aims to mitigate mis-calibration via knowledge distillation from larger models, especially in the field of decoder-only models which provides a new point of view to solve the problem.

2.2 Knowledge Distillation

Knowledge Distillation (KD) can be viewed as a transfer learning that allows a weak and small model (student model) to learn from a strong and large model (teacher model) in order to deduce the model size while preserving good performance. Studies over the past years have provided important information on how to distill encoder-only language models (Hinton et al., 2015), but KD on large decoder-only language models is still underexplored. Based on whether we can access prediction probability or not, 'we categorize existing distillation methods into two types.

2.2.1 Black-box Distillation

Black-box models refer to models that we are unable to access their weight and prediction logits such as ChatGPT (Ouyang et al., 2022), and PaLM (Chowdhery et al., 2022). Given an input, we are only able to get the next token without its probability distribution. Recent studies have made the attempt 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. Different from these methods, this study focus on models that we have access to the output probability distribution as it can provide richer information and lead to better performance.

2.2.2 White-box Distillation

White-box models mean the models are either fully open-sourced such as Llama (Touvron et al., 2023) 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 is typically optimized by a distribution match between teachers and students, potentially producing better small models given the more fine-grained signals (Gu et al., 2023). Our work is an extension of white-box distillation and focuses on how white-box distillation can be improved by re-calibrating the teacher signals.

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) involves keeping both the teacher model and the student model on the GPU simultaneously during training as shown in Figure 2(b). The advantage of this approach is that we can access the teacher's entire vocabulary distribution. However, the downside is that the teacher model occupies a significant amount of GPU mem-

ory, resulting in low optimization efficiency and slow training speed, which can hardly scale up to very large teacher models (i.e. $>$ 30B).

On the other hand, offline distillation typically involves generating distribution data from the teacher model beforehand. During optimization, only the student model is on the GPU. The drawback of this method is that we cannot store the probability of every token, as it would consume too much disk space. Instead, the top-k probabilities are kept. Aiming to provide efficient and practical distillation algorithms, our study focuses on offline distillation and keeps the top-5 probability, which has a trade-off between performance and resource as shown in Figure 2(c)(d).

3 Method

As shown in Figure 2(d), our approach can be mainly divided into four stages: *teacher building*, *efficient offline data generation*, *Re-calibration* and end with an illustration of *distribution matching*.

Teacher Building After receiving a train set D for the downstream task, we first use it to fine-tune a large teacher model by optimizing a normal language modeling loss:

$$Loss(y_{1:N}) = -\sum_{t=1}^{N} \log\ p(y_t|y_{<t})$$

where $y_1, y_2, ...y_N$ is a training token sequence.

Efficient Data Generation Given the train set D and fine-tuned teacher model, we first prepare the distillation data which contains top-5 probability in the offline setting. This is because for a large teacher model (e.g.>10B), it is inefficient to place both the teacher and student model on GPUs due to the their heavy memory consumption. In addition, retrieving the probability distribution in advance for each token entry may occupy large disk space. For example, given a 50,000-token vocabulary, retrieving the full probability from a dataset of 100,000 samples with an average length of 2,048 requires 120 TB storage, which is highly impractical. Given that the top-5 probability typically accounts for over 95% of the total probability in most cases, and our method is expected to be naturally extended to distillation from GPT-3 series, we choose only to generate the top-5 probability for further distillation.

Re-calibration After collecting the top-5 token probability from the teacher model, we first apply a re-calibrate on the probability distribution of the validation set to select an optimal smoothing coefficient c that results in the lowest ece. Then we re-calibrate the generated teacher probability on the training set for further distillation by:

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

In our setting, $i, j = 1, \ldots, 5$, representing the Top-5 token probability.

Distribution Match 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:

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

4 Experimental Setting

In this section, we first introduce the experiment setting which includes datasets and metrics (§ 4.1). The models (§ 4.2), baselines (§ 4.3), are presented in the following two subsections respectively. Implementation details can be checked in Appendix A.1.

4.1 Datasets and Evaluation Metrics

We conducted the experiments under two practical setups: (i) Direct training on individual downstream tasks and testing on the same task, so-called **In-Domain** which has been widely used to adapt language models to specific domains. (ii) General training on instruction-following tasks (Instruction tuning) and testing it on unseen downstream tasks, which is called **Out-of-Domain** setting. In such cases, we want to test the calibration ability and performance on a general-purpose model.

• **In-Domain:** We conduct experiments on two commonly used question answering tasks CommonsenseQA(CSQA) (Talmor et al., 2019) and BoolQ (Clark et al., 2019), respectively. We manually split 10% data to serve as a validation set and

then measure the calibration and performance on the corresponding test set.

• **Out-of-Domain:** Different from our in-domain setting, we first train the model on Alpaca (Taori et al., 2023), a commonly used instruction-following dataset. Then we measure the calibration and performance on CommonsenseQA and OpenBook QA(OBQA) (Mihaylov et al., 2018), which are particularly adopted to reflect the model's confidence level in terms of reasoning ability after training on general instruction-following data.

Metrics To measure calibration, we treat the free-text generation task as a classification task by restricting the model to generate only one token. We then obtain the highest probability choice over this token entry from a set of candidate choices (i.e. A/B/C/D) using $\arg\max_{i \in C} P(i)$, where C represents the set of candidates. We use the retrieved token probability as the predicted confidence, and its corresponding choice to calculate accuracy by comparing it to the ground truth. Finally, we compute the expected calibration error (ECE) as follows:

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

Here we consistently set the number of bins to 10. To measure performance, we check whether the prediction is an exact match of the ground truth, and compute the final accuracy accordingly.

4.2 Language Models

We conduct all experiments on three different model families: Llama-1, Llama-2 (Touvron et al., 2023) and OpenLlama (Geng and Liu, 2023) as they all have a wide range of model sizes from small to large, which make the distillation possible. In addition, these models have been widely used by the community and have shown strong ability on instruction following and reasoning tasks that make our results more reliable and useful. For Llama-1, we adopt Llama-1 7B as the student model and corresponding Llama-1 33B fine-tuned model to be the teacher. While for Llama-2, we choose 13B model as the teacher model since 34B model is not publicly available and 70B model requires significantly more training resources. For OpenLlama, we choose the largest 13B model in this model family as the teacher model.

4.3 Baselines

We set up five baseline methods and our proposed COD pipeline as follows:

- Pre-train: The original model without any tuning which has the same size as the student model. It typically has the lowest performance on downstream tasks.
- Teacher: The large model fine-tuned on D with hard labels.
- Fine-tune: The small model fine-tuned on D with hard labels, which has the same size as the student model.
- Distill $_{\text{7B w/o Re-calibration}}$: The distilled model by preserving the original teacher-generated probability distribution without re-calibration.
- Distill $_{\text{7B w/ Label Smoothing}}$: The distilled model by re-calibrating the original teacher-generated probability distribution using Label Smoothing.
- Distill $_{\text{7B w/ COD}}$: The distilled model by re-calibrating the original teacher-generated probability distribution using our proposed COD pipeline.

5 Experiment Results

To establish the importance of re-calibration during distillation and the superiority of COD pipeline over standard distillation and fine-tuning, we first compare the calibration ability of pre-training, fine-tuning, and distillation with and without re-calibration on two in-domain tasks. We show that, with COD, distillation can generally improve both the calibration and performance on downstream tasks compared with other distillation pipelines and standard fine-tuning. Finally, we extend the results to out-of-domain settings. The overall results are shown in Table 1.

5.1 In-Domain Results

As shown in the left part of Table 1, we evaluate the calibration and performance of several baselines and our COD pipeline on the CSQA and BoolQ test set. We conclude the following findings:

• **Fine-tuning lead to catastrophic miscalibration**: We observe that fine-tuned models generally exhibit worse calibration compared to pre-trained counterparts. For example, both the

	In-Domain				Out-of-Domain			
	Commonsense QA		BoolQ		Alpaca → Commonsense QA		Alpaca → OBQA	
	ECE ↓	Acc ↑	ECE ↓	Acc ↑	ECE ↓	Acc ↑	ECE ↓	Acc ↑
	LLAMA 1 : 33B → 7B							
Pre-train $_{7B}$	0.042	28.1%	0.425	61.3%	0.039	27.8%	0.04	27.2%
Teacher $_{33B}$	0.102	82.4%	0.077	89.7%	0.186	69.2%	0.202	64.4%
Fine-tune $_{7B}$	0.118	79.9%	0.065	82.5%	0.125	48.2%	0.219	43.4%
Distill $_{7B\ w/o\ Re\text{-}calibration}$	0.094	78.9%	**0.04**	85.3%	0.053	43.1%	0.181	39.8%
Distill $_{7B\ w/\ Label\ Smoothing}$	0.091	78.1%	0.190	85.3%	0.052	43.9%	0.19	37.6%
Distill $_{7B\ w/\ COD}$	**0.029**	**80.8%**	**0.04**	**85.7%**	**0.046**	**50.0%**	**0.071**	**47.2%**
COD to w/o Re-calibration	↑$_{6.5\%}$	↑$_{1.9\%}$	↑$_{0\%}$	↑$_{0.4\%}$	↑$_{0.7\%}$	↑$_{6.9\%}$	↑$_{11\%}$	↑$_{7.4\%}$
	LLAMA 2 : 13B* → 7B							
Pre-train $_{7B}$	0.1	36.6%	0.386	57.2%	0.1	36.6%	0.125	44.7%
Teacher $_{13B}$	0.12	81.6%	0.068	89.7%	0.208	65.7%	0.287	58.3%
Fine-tune $_{7B}$	0.14	76.8%	0.084	87.5%	0.212	50.0%	0.301	45.6%
Distill $_{7B\ w/o\ Re\text{-}calibration}$	0.109	80.0%	0.04	85.3%	0.077	50.9%	0.125	46.6%
Distill $_{7B\ w/\ Label\ Smoothing}$	0.103	**80.4%**	0.039	87.5%	0.075	51.1%	0.162	47.6%
Distill $_{w/\ COD}$	**0.063**	80.3%	**0.014**	**87.9%**	**0.055**	**51.4%**	**0.081**	**49.5%**
COD to w/o Re-calibration	↑$_{4.6\%}$	↑$_{0.3\%}$	↑$_{2.6\%}$	↑$_{2.6\%}$	↑$_{2.2\%}$	↑$_{0.5\%}$	↑$_{4.4\%}$	↑$_{2.9\%}$
	OPENLLAMA : 13B → 7B							
Pre-train $_{7B}$	0.075	20.8%	0.359	58.5%	0.075	20.8%	0.008	28.4%
Teacher $_{13B}$	0.132	78.5%	0.075	87.6	0.167	49.5%	0.134	50%
Fine-tune $_{7B}$	0.105	75.0%	0.036	81.5%	0.216	28.3%	0.161	30.4%
Distill $_{7B\ w/o\ Re\text{-}calibration}$	0.092	75.2%	0.062	83.8%	0.097	27.7%	0.137	29.8%
Distill $_{7B\ w/\ Label\ Smoothing}$	0.096	74.5%	0.033	83.3%	0.041	29.2%	0.142	29.8%
Distill $_{7B\ w/\ COD}$	**0.050**	**77.2%**	**0.027**	**84.7%**	**0.029**	**30.5%**	**0.082**	**30.8%**
COD to w/o Re-calibration	↑$_{4.2\%}$	↑$_{2.0\%}$	↑$_{3.5\%}$	↑$_{0.9\%}$	↑$_{6.8\%}$	↑$_{2.8\%}$	↑$_{5.5\%}$	↑$_{1.0\%}$

Table 1: The overall experimental results of calibration and performance on downstream tasks under both in-domain and out-of-domain setting. We compare pre-trained models, fine-tuned teacher and student models, and distilled models w/ or w/o re-calibration. The ↑ represents the larger the better while the ↓ means the smaller the better. **Bold** represents the best among fine-tuned and distilled student models. Gray represents the statistics presented are for reference only and should not be used for comparison purposes. The model sizes are all specified in the subscripts. *: We use 13B teacher model for Llama 2 family as its 34B version is still not publicly available and its 70B version requires significantly more resources.

fine-tuned student model and teacher have higher ece values than the pre-train 7B model on CSQA of three model settings.

It is also observed that pre-train models that deviate significantly from random guess performance tend to show larger mis-calibration. This is because these models have not been fine-tuned on the specific dataset and is supposed to produce random guess probability. When high accuracy is observed, there would be a mismatch between its prediction and true likelihood, leading to large mis-calibration rate.

• **Direct distillation brings bad calibration as well**: Furthermore, distilled models without re-calibration show varied calibration ability and performance. For in-domain tasks, the distilled Llama-1 and Llama-2 7B without re-calibration have ece values of 9.4% and 10.9% on CSQA, 4.0% and 4.0% on BoolQ respectively, a mis-calibration level similar to fine-tuned models. And distilled model of OpenLlama shows even worse calibration than fine-tuned models on BoolQ, indicating bad calibration ability of distillation without re-calibration under in-domain setting. While for performance, direct distillation 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 finding suggests that distillation without re-calibration does not consistently lead to good calibration and performance, demonstrating that matching mis-calibrated teacher distribution cannot result in a well-calibrated student model and verifying our motivation.

• **Re-calibration before distillation can greatly improve the calibration ability**: In contrast, distilled models with re-calibration, particularly our proposed COD pipeline, demonstrate significantly better calibration, achieving the lowest ece scores across all tasks, with an improvement up to 13% over the fine-tuning models of the same size. This suggests that COD effectively adjusts the model's confidence levels to match its prediction accuracy.

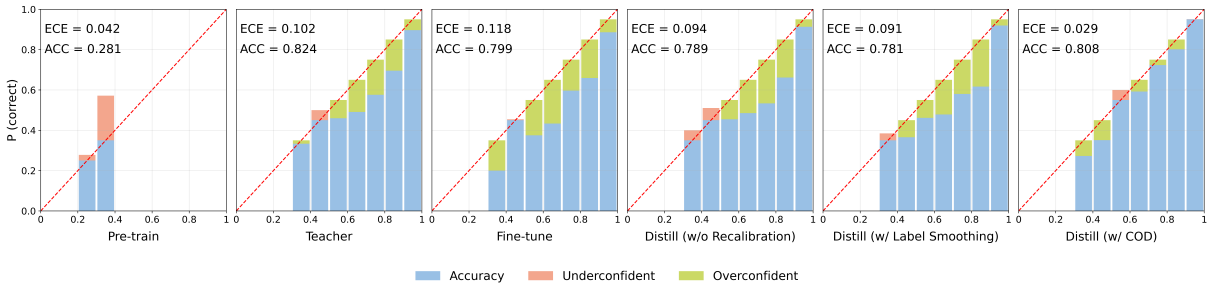


Figure 3: A reliability diagram showing mis-calibration comparison between pre-trained 7B, fine-tuned 33B (Teacher), fine-tuned 7B (Fine-tune) and distillation methods with or without re-calibration of Llama-1 family on Commonsense QA. Distill with COD is the rightmost figure. The red bar means accuracy higher than perfect calibration (under-confident) while the green bar means accuracy lower than perfect calibration (over-confident). The X-axis is 10 bins according to the model's confidence in each of the multiple choices for each question while the Y-axis is the accuracy within each bin.

In such cases, the model's overconfident issue is mitigated. Another baseline, distillation with label smoothing, also improves calibration but doesn't reach the efficacy levels of COD, as COD considers an optimal smoothing coefficient on the validation set, we further analyze the effectiveness of the optimal smoothing coefficient in Section 6.1.

On top of calibration, distilled models with re-calibration also show consistently improved performance on downstream tasks. Specifically, COD has an averaged improvement of 2.5% accuracy over non-calibrated baselines, indicating it not only makes well-calibrated confidence in its predictions but also has a more accurate prediction, underscoring the importance of re-calibration in distillation processes.

5.2 Out-of-Domain Results

The out-of-domain experiment statistics as shown in the right part of Table 1 indicate that the general trends observed in in-domain settings hold true in out-of-domain settings as well.

When models are fine-tuned on instruction following tasks and then applied to out-of-domain tasks, they also exhibit a poor level of calibration, indicated by higher ece values. This trend is consistent with the in-domain findings, which means that current general-purpose models suffer from great mis-calibration in many of their downstream tasks, making them unreliable to be deployed in real-world applications. For instance, the fine-tuned 7B model shows 3×-13× increased ece values in the out-of-domain tasks compared to the pre-trained models.

Distilled models without re-calibration show some improvement in calibration over fine-tuned models, but this improvement is not strong and consistent. For example, the distillation for Llama-1 and Llama-2 models without re-calibration does show reduced ece values compared to their fine-tuned counterparts on out-of-domain settings, however, these values are not low enough to indicate strong calibration. This is especially evident with the OpenLlama model, which, despite being distilled, displays worse calibration on BoolQ than fine-tuned models, highlighting the limitations of distillation without re-calibration in out-of-domain settings.

In contrast, COD pipeline consistently outperforms both the standard distillation and fine-tuning approaches in terms of calibration on out-of-domain tasks. They achieve the lowest ece values across all out-of-domain tasks across three model families, indicating that the re-calibration process within COD is effective even when applied to general instruction tuning and tested on unseen real-world tasks. This suggests that COD is a robust pipeline for solving mis-calibration issues and improving performance under real-world out-of-domain settings.

6 Additional Experiments and Analysis

6.1 Visualization on Calibration

In addition to metric-based analysis, we also draw a reliability diagram for better visualization and compare the mis-calibration of each model. As shown in Figure 3, each sub-plot contains a reliability diagram for pre-trained Llama-1-7B, fine-tuned Llama-1-33B, fine-tuned Llama-1-7B, and three distilled methods from left to right. A perfectly calibrated model would have a straight diagonal line from the bottom left to the top right of such a diagram, indicating that confidence level is exactly consistent with actual accuracy. The reliability dia-

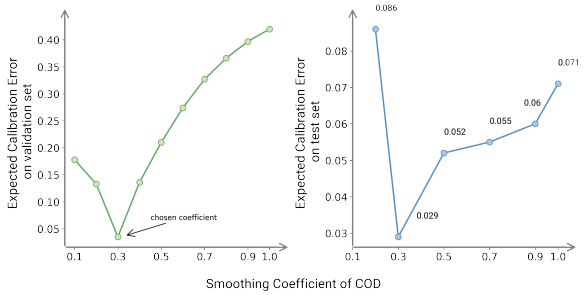


Figure 4: Left shows the comparison of different smoothing coefficient on the validation set, while the right part demonstrates its corresponding calibration effect on the test set.

grams are divided into 10 bins based on the model's confidence. The bars represent the accuracy within each bin, and the colors indicate whether the model is under-confident (red) or over-confident (green) within each bin. The class-wise expected calibration error at the top of each plot provides a quantitative measure of calibration, with a lower score indicating better calibration. The figure indicates that the pre-trained model exhibits bad calibration as it deviates from the perfect calibration line a lot. Since it has not been fine-tuned for this specific task, it predicts probabilities only range from 0.2 to 0.4. After fine-tuning, the model shows catastrophic mis-calibration where it introduced a degree of over-confidence and under-confidence in the model's predictions. The fine-tuned model's bars now frequently rise below the perfect calibration line, indicating that it often predicts with higher confidence than is warranted by its actual accuracy. However, we also observe that the large teacher model (33B) is more calibrated than the small model (7B) and the direct distilled model can reach slightly better calibration. While for distillation with COD pipeline, the under-confident and over-confident are noticeably less than fine-tuned models and distilled models without re-calibration, as evidenced by a smaller area of red bar and green bar and its proximity to the perfect calibration line.

6.2 Effectiveness of Smoothing Coefficient in COD

As discussed in the COD pipeline, we pre-select an smoothing coefficient that reach a lowest ece score based on the validation set as shown in the left part of Figure 4. We first divide the interval from 0 to 1 by a step of 0.1, as a coefficient of 1 already compressed the token probability a lot and the therefore coefficient of 0.1 will even enlarge the probability of over-confident tokens. We further locate a smaller interval that contains potential optimal value and use a smaller step of 0.02 to find the best smoothing coefficient. We further compare COD distillation with the selected optimal smoothing coefficient and other different smoothing coefficient as shown in the right part of Figure 4. COD with optimal smoothing coefficient do outperform those with other levels of smoothing coefficient with a large margin, indicating the effectiveness of selecting such optimal smoothing coefficient.

6.3 Comparison to Post-calibration after Training

We also compare our COD method with some post-calibration techniques such as temperature scaling after direct distillation and fine-tuning. These post-calibration are often regard as poorer performance than consider re-calibration during training and will compress the final prediction probability a lot which make the confidence non-differentiable. For example, direct distilled OpenLlama on Alpaca after post temprature scaling still have a 30% ece higher than its COD counterparts and with its highest confidence reduced to 60%.

7 Conclusion

In conclusion, this study presents a novel Offline knowledge Distillation with Re-calibration (COD) approach for large language models (LLMs), emphasizing the importance of re-calibration in knowledge distillation. The COD pipeline effectively addresses the issue of mis-calibration in distilled models, consistently leading to more reliable and accurate student models. Through comprehensive experiments, it's shown that re-calibration prior to distillation significantly improves model reliability and performance in both in-domain and out-of-domain settings. The research contributes to the understanding of the mis-calibration issue in current LLM tuning methods and offers a robust method for enhancing their performance and calibration ability, proving particularly useful in scenarios involving large teacher models.

8 Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. We address the critical issue of catastrophic mis-calibration in current training pipelines (supervised fine-tuning and knowledge distillation) and propose a pipeline to obtain a more reliable model. There are many po-

tential societal consequences of our work, none which we feel must be specifically highlighted here.

9 Limitations

It is shown that our Calibrated Offline Knowledge Distillation (COD) demonstrates superior calibration ability and performance over direct distillation and standard fine-tuning methods. However, despite these exciting results, there are still some limitations to 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 by employing large teacher model such as 70B can lead to better calibration ability and performance since large model learn a better distribution. However, we are unable to explore how very large teacher perform due to resource limitation.

Top-K Chosen in Offline Distillation: Another limitation of this work is that it does not provide a rigorous study on how many token probability 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 following reasons: (1) We suppose the top-5 token probability already contain most of the information (i.e. sum of top-5 probability is close to 1) for the whole distribution and top-5 will not consume tremendous disk space (2) Current strong gray-box models like *text-davinci-003* from OpenAI can only return the top-5 probability for each token entry, so that our method can be extend to the data generated by these models.

However, how many 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.

Yangyi Chen, Lifan Yuan, Ganqu Cui, Zhiyuan Liu, and Heng Ji. 2023. A close look into the calibration of pre-trained language models.

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.

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.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531.

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.

S. Kullback and R. A. Leibler. 1951. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86.

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.

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

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

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.

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. 2023. Llama: Open and efficient foundation language 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.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

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.

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.

	Fine-tuned	Distillation w/o re-calibration	Distillation w/ COD
Question	Which city is farther north, Oslo or Helsinki?		
Correct Answer	Helsinki		
Generated Confidence	Oslo is farther north than Helsinki. 0.73 → over-confident	Oslo is farther north than Helsinki. 0.79 → over-confident	Oslo is farther north than Helsinki. 0.54
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. 0.78

Table 2: A case study on how fine-tuned model and distilled model without re-calibration tend to over-confident on the wrong answer with high confidence. While distillation with COD though output a wrong answer but it produce 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 and cosine annealing scheduler with a warmup 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. 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. 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.

B Additional Analysis

B.1 Case Study

We further conduct a case study to see whether re-calibration indeed helps mitigate mis-calibration in real-world question answering. As shown in Table 2, we ask the models that use three different tuning methods on Alpaca a 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 fine-tuned models and distillation w/o re-calibration give high confidence in the wrong answer, which is far from satisfactory for real-world settings, especially when additional post-processing procedures were expected to be applied to filter wrong answer by identifying unconfident responses. In comparison, distillation with COD greatly mitigates this mis-calibration by producing a confidence around 50% which indicate the model is not sure about the generated answer, allowing systems to filter those undesirable answers by a hard confidence threshold.