MiCoTA: Bridging the Learnability Gap with Intermediate CoT and Teacher Assistants

Anonymous authors
Paper under double-blind review

Abstract

Large language models (LLMs) excel at reasoning tasks requiring long thought sequences for planning, reflection, and refinement. However, their substantial model size and high computational demands are impractical for widespread deployment. Yet, small language models (SLMs) often struggle to learn long-form CoT reasoning due to their limited capacity, a phenomenon we refer to as the "SLMs Learnability Gap". To address this, we introduce **Mi**d-**Co**T **T**eacher **A**ssistant Distillation (MiCoTA), a framework for improving long CoT distillation for SLMs. MiCoTA employs intermediate-sized models as teacher assistants and utilizes intermediate-length CoT sequences to bridge both the capacity and reasoning length gaps. Our experiments on downstream tasks demonstrate that although SLMs distilled from large teachers can perform poorly, by applying MiCoTA, they achieve significant improvements in reasoning performance. Specifically, Qwen2.5-7B-Instruct and Qwen2.5-3B-Instruct achieve an improvement of 3.47 and 3.93 respectively on average score on AIME2024, AMC, Olympiad, MATH-500 and GSM8K benchmarks. To better understand the mechanism behind MiCoTA, we perform a quantitative experiment demonstrating that our method produces data more closely aligned with base SLM distributions. Our insights pave the way for future research into long-CoT data distillation for SLMs.

1 Introduction

Reasoning has long been regarded as one of the most challenging capabilities to instill in Large Language Models (LLMs). The advent of Chain-of-Thought (CoT) prompting (Wei et al., 2022) marked a significant milestone, revealing the emergent reasoning abilities of large-scale models when prompted to generate step-by-step thought processes. Building on this, a lot of work (Hao et al., 2023; Yao et al., 2023; Zelikman et al., 2022; Qi et al., 2024; Wan et al., 2024) has been developed to improve LLM reasoning by scaling inference time compute. The OpenAI o1 series (OpenAI, 2024a;b) became the breaking point where they enhanced effective test-time computation by encouraging LLMs to explore possible solutions and generating longer thinking sequences during inference. Following o1, other proprietary LLMs (DeepSeek-AI et al., 2025; Seed et al., 2025) also demonstrate remarkable performance on tasks requiring intricate reasoning and decision-making. However, the substantial model size and high computational demands of these state-of-the-art reasoners make them impractical for widespread deployment, particularly in resource-constrained environments. This necessitates the development of smaller, more efficient models that retain strong reasoning capability.

Knowledge distillation (Hinton et al., 2015) is one of the most promising strategy for transferring the capabilities of large teacher models to smaller student models. In the context of long CoT, most works have focused on distilling CoT trajectories, collecting vast datasets of complex queries paired with detailed reasoning steps generated by powerful LLMs (Team, 2025a; Face, 2025; Ye et al., 2025). The primary goal is to equip small language models (SLMs) with the ability to perform long-form reasoning, thereby bridging the performance gap between teacher models and student models. Several successful attempts (DeepSeek-AI et al., 2025; Team, 2024b; Ye et al., 2025) achieved promising results at a 32 billion parameters scope.

Despite progress, SLMs still struggle to learn long-CoT reasoning effectively (Li et al., 2025; Yin et al., 2025), which could be attributed to their limited capacity. We refer to this phenomenon as "SLMs Learnability

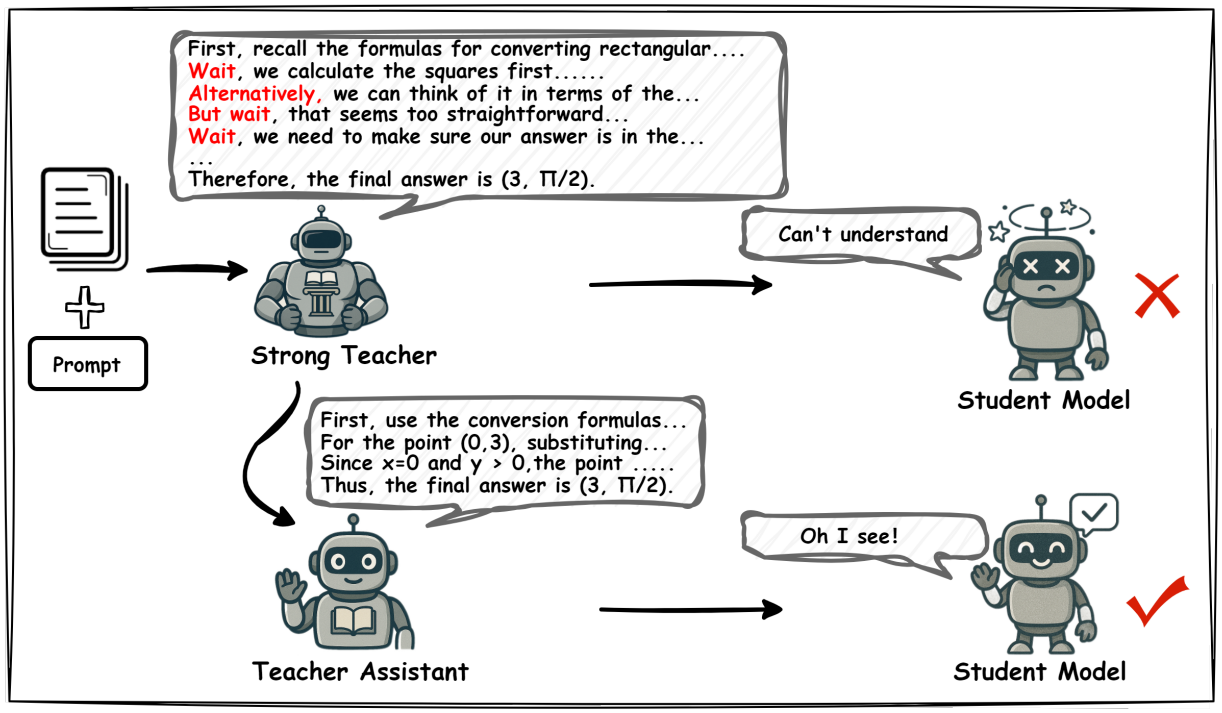

Figure 1: This figure illustrates the procedure and core concept of MiCoTA. While long CoT data distilled from a strong teacher is too lengthy for the student model to effectively learn, the half-length CoT data generated by an intermediate-sized Teacher Assistant model is more accessible and easier for the student model to benefit from.

Gap" where the performance of SLMs degrades when trained on long CoT sequences distilled from large teacher models. In this paper, we conduct an in-depth investigation to unveil and better understand the underlying causes of this phenomenon. We propose a multifaceted approach called **Mi**d-**CoT** **T**eacher **A**ssistant Distillation (MiCoTA), illustrated in Figure 1. MiCoTA leverages an intermediate-sized model as a teacher assistant (TA) model and distills intermediate-lengthed CoT data to train the student model.

To begin with, we conducted pivotal studies to unveil this counterintuitive phenomenon. Our experiments with models trained on distillation data from teachers of varying sizes revealed that not only did models trained with larger teachers perform worse, but even those distilled from intermediate-sized teachers could not match the performance of the original base models (Yang et al., 2024). This finding led us to explore the complementary dimension of the problem: rather than focusing solely on model size, we investigated the impact of generating intermediate-length CoT reasoning data to bridge the gap between the comprehensive reasoning of large teachers and the learning capacity of smaller students. Specifically, inspired by previous work on long-to-short reasoning (Wu et al., 2025), we found that simply merging the long-CoT data trained model with its base model resulted in a new model that produced reasoning sequences approximately half the length of the original long-CoT output without a performance drop. We used these intermediate-length CoT data from the merged TA model to train the student model. This method addresses the SLMs Learnability Gap by mitigating both the capacity gap between the student and teacher models, as well as the length gap by learning from intermediate-length CoT, allowing smaller models to better benefit from large reasoning LLMs.

We apply our method to Qwen-series SLMs (Yang et al., 2024) spanning 1.5B, 3B and 7B parameters and evaluate them across various benchmarks (Hendrycks et al., 2021; He et al., 2024; Cobbe et al., 2021). Our experiments demonstrate that MiCoTA has significantly improved the reasoning performance of SLMs comparing to those directly trained on distilled data from the teacher model. Specifically, applying MiCoTA

results in a 9.98% improvement in reasoning performance for Qwen2.5-3B-Instruct over its base model and a 35.6% increase over its version distilled from a larger teacher model. We also conduct ablation studies to thoroughly analyze the effectiveness of our method.

Our main contributions are summarized as follows:

1. We propose a novel approach, Mid-CoT Teacher Assistant Distillation (MiCoTA), which effectively mitigates the SLMs Learnability Gap and achieves significant improvements in reasoning performance for SLMs.

2. We conduct thorough extensive studies to analyze the impact of various components of the MiCoTA approach. These studies confirm the effectiveness of our method and provide insights into the factors that contribute to improved reasoning performance for smaller models.

2 Related Work

2.1 Knowledge Distillation for LLM Reasoning

Knowledge distillation (KD) has long been a fundamental technique for transferring knowledge from large, powerful teacher models to student models (Hinton et al., 2015; Sanh et al., 2019; Sun et al., 2019; Jiao et al., 2020; Wang et al., 2020; Zhou et al., 2022; Xu et al., 2021). Early works in KD focused on matching the output logits of the teacher and student (Hinton et al., 2015; Beyer et al., 2022; Tang et al., 2019; Park et al., 2021; Zhao et al., 2022; Sanh et al., 2019), while subsequent research extended this idea to internal representations such as attention maps and hidden states (Romero et al., 2015; Sun et al., 2019; Zagoruyko & Komodakis, 2016; Wang et al., 2022; Jiao et al., 2020; Wang et al., 2020; Xu et al., 2020).

Beyond logits distillation (Hinton et al., 2015) and internal feature distillation (Wang et al., 2020; 2022), sequence-level distillation (Kim & Rush, 2016) has also gained traction since the rise of large language models (LLMs). Previous works (Wang et al., 2023; Xu et al., 2023; Taori et al., 2023; Chiang et al., 2023) explored distilling knowledge from GPT-4 by generating synthetic instruction data and leveraging teacher-generated responses for student training.

Recently, Chain-of-Thought (CoT) data distillation has emerged as a promising direction to enhance the reasoning capabilities of student models (Team, 2025a; Face, 2025; Wen et al., 2025; Li et al., 2025). By training student models on reasoning traces generated by strong LLMs, these methods aim to transfer complex multi-step reasoning skills. However, directly distilling long CoT traces from strong teachers often overwhelms smaller models, leading to suboptimal performance (Li et al., 2025).

2.2 The Learnability Gap

The learnability gap—the challenge arising from large discrepancies in capacity or architecture between teacher and student models—has been recognized as a long-standing issue in knowledge distillation (Mirzadeh et al., 2020; Zhang et al., 2023b). To address this, the teacher assistant (TA) paradigm introduces an intermediate model that bridges the gap between the strong teacher and the small student (Mirzadeh et al., 2020; Son et al., 2021; Zhang et al., 2023a). TAs have been implemented across various settings, including differences in model size (Mirzadeh et al., 2020), architecture, and submodules, to facilitate more effective knowledge transfer.

Despite the success of TAs in standard KD scenarios, their application in long-CoT distillation remains underexplored. Few studies have investigated the use of TAs to address the length aspect of the gap curse in CoT distillation. Recent work (Li et al., 2025) proposed mixing long-CoT and short-CoT data to ease the learning burden on student models. On the other hand, our approach proposes directly generating intermediate-length CoT data with a merged TA model (Goddard et al., 2024), effectively filling the gap and enabling small language models to benefit from rich teacher reasoning without being overwhelmed.

3 Methodology

In this section, we propose to use an intermediate-sized model as a teacher assistant (TA) to bridge the gap between teacher models and student models. The key idea is to train an LLM with both model size and CoT length set to approximately half of the teacher. We pose that not only the curse of capacity gap happened on the size of the model, but also the length of the reasoning path. Therefore this dual "half-size, half-length" design ensures that the teacher assistant is not only more accessible for the student model to learn from in terms of capacity, but also provides reasoning demonstrations that are less overwhelming in length, thereby facilitating more effective knowledge transfer.

As illustrated in Figure 1, our Teacher Assistant is first trained using long CoT generated by the Strong Teacher. Afterward, we use the model merging method to merge the Teacher Assistant before and after fine-tuning to arrive at a final, Mid-CoT Teacher Assistant. This Mid-CoT Teacher Assistant is then employed to generate CoT that the Student Models can utilize for more effective training.

3.1 Reasoning Data Generation

We employ **R1-Distill-Qwen-32B** (DeepSeek-AI et al., 2025) as our strong teacher to generate exemplar long CoT. We craft prompts that explicitly request comprehensive reasoning steps. This ensures the generated CoT captures intermediate reasoning processes rather than just final answers. For each prompt, the strong teacher produces a detailed long CoT trace, including intermediate reasoning steps and the final answer. After data generation, we further exclude any entries with incorrect answers or those that exceed the maximum length to ensure the quality of the dataset. We compile these traces into a dataset:

$$D_{strong} = \left\{ \left(x^{(i)}, CoT_{strong}^{(i)} \right) \mid A_i = \text{true and } L_i \leq L_{\max} \right\}_{i=1}^{N}, \tag{1}$$

where A_i indicates whether the answer is correct, L_i represents the length of the generated CoT, $L_{\max}$ is the maximum allowable length, $x^{(i)}$ is the prompt, and $CoT_{strong}^{(i)}$ is the corresponding long CoT.

3.2 Teacher Assistant Creation

Training We select a medium-scale model with parameters less than those of the Strong Teacher but greater than those of the Student Model. We then fine-tune this model on D_{strong}, resulting in a model that learns to approximate the reasoning style of the Strong Teacher.

Model Merging Previous research (Wu et al., 2025) has shown that model merging effectively facilitates long-to-short reasoning by combining the quick-thinking abilities of System 1 models with the analytical reasoning of System 2 models. This approach directly manipulates model parameters without additional training. Figure 2 illustrates that model merging can reduce the response lengths of System 2 models by approximately half while maintaining performance.

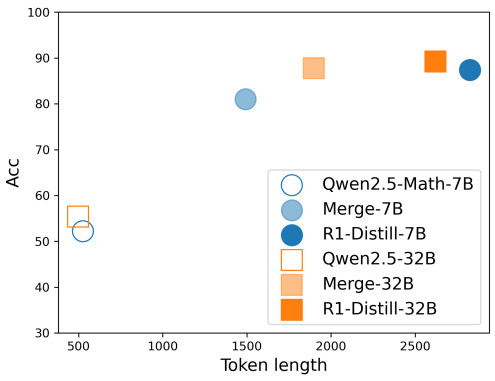

Figure 2: The token length is reduced to about half of the System 2 models by model merging.

To further reduce response length while balancing performance, we merge the pre-fine-tuned and post-fine-tuned versions of the Teacher Assistant model to achieve an intermediate response length. We utilize the *Dare* (Yu et al., 2024) algorithm to merge the Teacher Assistant before and after fine-tuning. *Dare* reduces interference between different models through sparsification of task-specific delta vectors. It utilizes random pruning in conjunction with a rescaling technique to preserve the original models' performance (Goddard et al., 2024).[1] In our

[1]Refer to `https://github.com/arcee-ai/mergekit` for the model merging codebase.

Table 1: Performance gap (Δ) between models trained on distillation CoT data and the original base models.

Student Models	QwQ-32B	Avg	Δ	R1-Distill-Qwen-32B	Avg	Δ
14B-Instruct	+Teacher	55.84	+3.66	+Teacher	54.90	+2.10
1.5B-Instruct	+Teacher	23.29	−6.97	+Teacher	21.25	−9.01
	+TA	25.16	−5.10	+TA	24.59	−5.67
3B-Instruct	+Teacher	32.00	−7.36	+Teacher	31.92	−7.44
	+TA	34.63	−4.73	+TA	33.08	−6.28

experiments, we combine *Dare* pruning with the TIES sign consensus mechanism to optimize the model merging process. After model merging, we obtain Mid-CoT Teacher Assistant, an intermediate state in both model size and inference length that inherits the complementary strengths of both models without additional supervised fine-tuning.

3.3 Student Model Training

We prompt the Mid-CoT Teacher Assistant to produce a synthetic medium-length CoT dataset $D_{MiCoTA} = \left\{(x^{(i)}, CoT^{(i)}_{MiCoTA})\right\}_{i=1}^{N}$. Unlike the original long CoT from the Strong Teacher, these new CoT sequences are shorter and presumably more accessible for the Student Models to learn. Using D_{MiCoTA}, we fine-tune the Student Models with standard supervised fine-tuning (SFT).

4 Experiments

In this section, we give a detailed introduction to our pivotal study, experimental setup, main results, ablation analysis, etc.

4.1 Pivotal Study

This section evaluates the performance of student models trained on Strong Teacher CoT and Intermediate-sized Teacher CoT. We choose QwQ-32B (Team, 2025b) and R1-Distill-Qwen-32B (DeepSeek-AI et al., 2025) as our Strong Teacher models to generate Teacher CoT responses. The Qwen2.5-14B-Instruct model, trained on CoT data generated by Strong Teachers, serves as the Intermediate-sized Teacher to produce TA CoT. Student models include three parameter-scale models: Qwen2.5-1.5B-Instruct, Qwen2.5-3B-Instruct, and Qwen2.5-14B-Instruct. To quantify the impact of different distillation data on student models, we define a performance difference metric Δ to measure the performance gap between trained models and their original baseline models, defined as:

$$\Delta = P_{distilled} - P_{base}. \tag{2}$$

Table 1 provides the results. A negative (positive) Δ indicates that the performance of models trained on CoT data is worse (better) than that of the original base models. The results indicate a nuanced relationship between student model size and performance when trained on distillation data from the Strong Teacher. Specifically, the larger student model, Qwen2.5-14B-Instruct, demonstrates a clear advantage when trained using the Strong Teacher CoT, achieving a positive performance gain. Conversely, smaller models (1.5B-Instruct and 3B-Instruct) experience negative performance changes when trained on the same CoT data, suggesting they may struggle to effectively learn from the longer CoT paradigm. Subsequently, we utilize the well-trained Qwen2.5-14B-Instruct as our Intermediate-sized Teacher Assistant model to generate CoT data for training the smaller student models. The results show that the Intermediate-Sized Teacher Assistant data provides improvements for the smaller models, but still cannot match the performance of the original base model.

4.2 Experimental Setup

Datasets We utilize a prompt set of 12K items from the MATH dataset (Lightman et al., 2023), which includes seven math subjects: advanced calculus, geometry, and linear algebra. For data generation, we apply greedy decoding with a maximum token limit of 16K. During this process, we exclude any entries that exceed the maximum length. By pairing math problem instructions with corresponding solutions from teacher models, we create problem-solution pairs for fine-tuning the student models.

Models We employ R1-Distill-Qwen-32B (DeepSeek-AI et al., 2025) as our strong teacher model. We utilize the Qwen2.5-14B-Instruct as our teacher assistant model. For the student models, we evaluate three SLMs from the Qwen family, which cover a range of parameter scales: Qwen2.5-7B-Instruct, Qwen2.5-3B-Instruct, and Qwen2.5-1.5B-Instruct.

Training All fine-tuning experiments were performed using the LLaMA-Factory framework (Zheng et al., 2024) on a server equipped with eight NVIDIA A800-SXM4-80GB GPUs. Detailed hyperparameters and information about the training setting are provided in Appendix A.

Evaluation We evaluate performance on the following benchmarks: AIME 2024, AMC 2023, OlympiadBench (He et al., 2024), MATH 500 (Hendrycks et al., 2021), and GSM8K (Cobbe et al., 2021). To make our empirical results more reproducible, we adopt greedy decoding during inference. All models are assessed in a zero-shot setting. The maximum length for the evaluated models is set to 16K tokens. Following Gao et al. (2024), we only adopt a rule-based method to extract the final answer out of model's generation and measure the exact-match accuracy[2]. We do not apply LLM-as-judge based evaluation in our experiments.

Baselines We compare MiCoTA against three baseline settings:

- **Instruct**: SLMs without trained on long-CoT data.
- **Strong Teacher CoT**: The SLMs trained on Long-CoT data generated by the Strong Teacher.
- **Mix-Long** (Li et al., 2025): Fine-tuning on data mixed with long CoT and short CoT data, where long CoT is generated from QwQ-32B and short CoT is generated from Qwen2.5-32B-Instruct.

Model Setting

- **Half-size CoT**: The SLMs trained on long-CoT data distilled from Intermediate-sized Teacher Assistant.
- **MiCoTA**: The SLMs trained on intermediate-length CoT data distilled from merged Teacher Assistant.

4.3 Main Result

Table 2 presents a comprehensive evaluation of model performance across different benchmarks. From the table, we can observe a performance decline in the student models after training with the Strong Teacher CoT. This finding is consistent with previous studies (Li et al., 2025; Wu et al., 2025), which suggest that smaller models often struggle to effectively learn from stronger teacher models. For instance, after training on the Strong Teacher CoT, the Qwen2.5-7B model exhibits a decrease in its average score compared to the instruct models without any training. When trained with the Mix-Long method, the model's performance improves. In contrast, the student models trained using the Mid-CoT Teacher Assistant outputs exhibit improvements. For example, the Qwen2.5-7B model achieves an average score of 49.36, surpassing all baseline methods. Similarly, the Qwen2.5-3B student model shows enhanced performance, with an average score of 43.29 when trained with the MiCoTA. The Qwen2.5-1.5B model also benefits from our approach, with its average score

[2]Refer to `https://github.com/EleutherAI/lm-evaluation-harness` for the evaluation codebase.

Table 2: Performance comparison across various benchmarks; the highest score is bolded, and the second highest score is underlined.

Models	Method	AIME	AMC	Olympiad	MATH-500	GSM8K	Average
Strong Teacher							
R1-Distill-Qwen-32B	–	60.00	90.00	41.92	86.00	86.27	72.83
(Merged) Teacher Assistant							
Qwen2.5-14B	–	23.33	62.50	31.55	77.40	87.26	56.40
Student Models							
Qwen2.5-7B	Instruct	10.00	40.00	24.44	**72.60**	82.41	45.89
	Strong Teacher CoT	6.66	27.50	20.00	64.20	84.38	40.54
	Mix-Long*	10.00	37.50	24.44	68.40	**88.85**	45.83
	MiCoTA	**13.33**	**52.50**	**25.62**	70.40	84.98	**49.36**
Qwen2.5-3B	Instruct	3.33	35.00	18.07	**62.80**	77.63	39.36
	Strong Teacher CoT	3.33	22.50	12.59	46.20	74.98	31.92
	Mix-Long*	10.00	37.50	19.11	60.20	**81.50**	41.66
	MiCoTA	**13.33**	**40.00**	**20.59**	61.60	80.97	**43.29**
Qwen2.5-1.5B	Instruct	**3.33**	22.50	12.44	45.20	67.85	30.26
	Strong Teacher CoT	0.00	7.50	7.85	33.00	57.92	21.25
	Mix-Long*	**3.33**	25.00	11.70	50.80	**71.65**	32.49
	MiCoTA	**3.33**	**27.50**	**13.03**	**51.00**	70.02	**33.01**

* The results presented here are based on our reproduction of the method.

Table 3: Ablation studies on MiCoTA with Qwen2.5-7B-Instruct, Qwen2.5-3B-Instruct, and Qwen2.5-1.5B-Instruct. The highest score is bolded, and the second highest score is underlined.

Models	Distillation Method	AIME	AMC	Olympiad	MATH-500	GSM8K	Average
Qwen2.5-7B	Strong Teacher CoT	6.66	27.50	20.00	64.20	84.38	40.54
	Half-size CoT	10.00	42.50	19.70	66.20	**86.27**	44.93
	MiCoTA	**13.33**	**52.50**	**25.62**	**70.40**	84.98	**49.36**
Qwen2.5-3B	Strong Teacher CoT	3.33	22.50	12.59	46.20	74.98	31.92
	Half-size CoT	3.33	22.50	12.30	50.40	76.88	33.08
	MiCoTA	**13.33**	**40.00**	**20.59**	**61.60**	**80.97**	**43.29**
Qwen2.5-1.5B	Strong Teacher CoT	0.00	7.50	7.85	33.00	57.92	21.25
	Half-size CoT	0.00	22.50	7.55	33.80	59.13	24.59
	MiCoTA	**3.33**	**27.50**	**13.03**	**51.00**	**70.20**	**33.01**

increasing to 33.01. We also extend our experiments to LLaMA family models in Table 8 and provide detailed LLM-judged benchmark scores in Table 9 (Appendix B). These findings collectively demonstrate that our method, which leverages the expertise of the Teacher Assistant for fine-tuning student models, effectively enhances their overall performance.

4.4 Ablation Studies

To validate the impact of our design choices, we conducted ablation studies on Qwen-Instruct models of varying sizes. Table 3 presents the results of different distillation methods, which include Strong Teacher CoT, Half-size CoT, and MiCoTA. These denote models trained on CoT data generated by the Strong Teacher, CoT data produced by the Intermediate-sized Teacher, and CoT data generated by our Mid-CoT Teacher Assistant, respectively. The results illustrated in Table 3 indicate that the student models trained on Half-size CoT outperform those trained on Strong Teacher CoT, highlighting its effectiveness as an intermediate-sized teacher model. Implementing an intermediate-sized teacher for small language models (SLMs) provides adequate guidance without overwhelming them. Notably, our MiCoTA approach shows a clear upward trend in performance across all model sizes, consistently surpassing the previous methods and achieving the highest scores among all configurations tested. Our MiCoTA strikes a moderate balance in

both model scale and response length: it preserves sufficient reasoning depth without causing information overload, thereby enabling SLMs to internalize complex multi-step reasoning patterns more effectively. We provide a case study of Half-size CoT and MiCoTA output in Appendix C. These results confirm the effectiveness of our proposed MiCoTA approach.

To verify the sensitivity of the results to the size of the Teacher Assistant (TA), we merge DeepSeek-R1-Distill-Qwen-32B and Qwen2.5-32B-Instruct as the half-length TA to train Qwen2.5-7B, Qwen2.5-3B, and Qwen2.5-1.5B student models. The model-merge strategy and all student-model training hyper-parameters are identical to those used when Qwen2.5-14B-Instruct served as the half-length teacher assistant. As shown in Table 4, Merged 32B teacher (half-length CoT) yields best performance for 7B student. For smaller models (3B/1.5B), 14B TA proves more effective. These results strengthen our original claims and provide more comprehensive evidence for MiCoTA's effectiveness and general applicability. The detailed results regarding sensitivity to the merge method and merge ratio are provided in Table 10 and Table 11 of Appendix B.

Table 4: Performance of Student Models Trained with Merged 32B Models as Half-Length Teacher Assistants.

	AIME	AMC	Olympiad	MATH-500	GSM8K	Average	Merged 14B as TA avg
Qwen2.5-7B	16.66	50.00	25.18	73.00	86.20	50.21	49.36 (-0.85)
Qwen2.5-3B	3.33	35.00	19.11	61.40	81.57	40.10	43.29 (+2.19)
Qwen2.5-1.5B	6.66	27.50	13.33	47.20	68.15	32.57	33.01 (+0.44)

Table 5: The results of adaptability of models on different CoT data.

Method \ Model	Qwen2.5-32B	Qwen2.5-14B	Qwen2.5-7B	Qwen2.5-3B	Qwen2.5-1.5B
Strong Teacher CoT	0.17	0.18	0.20	0.26	0.21
Half-size CoT	0.17	0.17	0.19	0.24	0.20
Mix-Long	0.13	0.14	0.14	0.15	0.16
MiCoTA	**0.11**	**0.11**	**0.13**	**0.13**	**0.14**

4.5 Adaptability of Models on Different CoT Data

We adopt the adapted Bits Per Character (BPC) metric from (Zhu et al., 2025), which is a variation of perplexity that eliminates length differences, to evaluate the model's adaptability with the text. The calculation is as follows:

$$BPC(T) = \frac{-\sum_{i=1}^{N} \log p(w_i|w_1, \cdots, w_{i-1})}{\text{len-utf-8}(T)} \tag{3}$$

Here, T denotes the text under analysis, N is the number of tokens in T, and len-utf-8(T) indicates the length of T in UTF-8 encoding, measured in characters. The tokens w_i are segments of text. By employing this metric, we are able to assess the models' orientations toward various distributions of CoT data and identify potential gaps in their learning processes. For this evaluation, we selected the models Qwen2.5-32B-Instruct, Qwen2.5-14B-Instruct, Qwen2.5-7B-Instruct, Qwen2.5-3B-Instruct, and Qwen2.5-1.5B-Instruct to analyze their adaptability on the Strong Teacher CoT data, Half-size CoT data, Mix-Long data, and the proposed MiCoTA data. The results are summarized in Table 5, which illustrates the BPC values across different models and data types. For the Strong Teacher CoT data, the smaller models exhibit relatively high BPC values, reflecting a larger distribution gap between the model predictions and the actual data distribution. This discrepancy may contribute to their difficulties in effectively learning from the Strong Teacher CoT data. In contrast, the BPC values for the Half-size CoT data are slightly lower than those for the Strong Teacher CoT data, suggesting a reduced distribution gap. This reduction indicates that the smaller models can adapt slightly better from the intermediate-sized teacher. Notably, for the MiCoTA

data, we observe considerably lower BPC values across all models compared to the Strong Teacher CoT, Half-size CoT, and Mix-Long data. This could suggest that the MiCoTA CoT data more closely matches their inherent distribution, thereby facilitating a more effective learning process. These findings validate the effectiveness of our proposed method.

4.6 Length Analysis

We further explore response lengths of student models under three configurations: the Instruct model, the one trained with Strong Teacher CoT data, and the one trained with MiCoTA CoT data. As shown in Figure 3, response lengths differ significantly across models and configurations. Across all sizes, the Instruct model—designed for concise, targeted answers—produces shorter responses. By contrast, the Strong Teacher CoT-trained version, which emphasizes chain-of-thought (CoT) reasoning, generates longer outputs via step-by-step analysis and elaboration. The version trained with MiCoTA, meanwhile, finds a balance between brevity and comprehensiveness, yielding outputs of medium length as it incorporates some elements of chain-of-thought reasoning without excessive elaboration. Additionally, response lengths generally increase as model size decreases. Smaller models, with limited parameter capacity, may need more verbose explanations to cover all reasoning steps; larger models integrate information more efficiently, delivering comprehensive yet succinct responses, unlike smaller ones that often rely on detailed descriptions for coherence and completeness. We further analyze **top 30** high-entropy tokens (Wang et al., 2025) in the reasoning responses of **DeepSeek-R1-Distill-Qwen-32B** and **MiCoTA** in Table 6. These tokens are considered strongly associated with successful model reasoning. The relative ranking of word choices reveals distinct reasoning styles: 1) **MiCoTA** prioritizes concise, imperative terms (e.g., "try," "consider") and formal logic markers (e.g., "by," "given"), reflecting a **structured, theorem-like** approach. 2) **DeepSeek** exhibits self-correction (e.g., "wait," "perhaps," "actually") and speculative language (e.g., "maybe," "think"), suggesting **exploratory, iterative** reasoning.

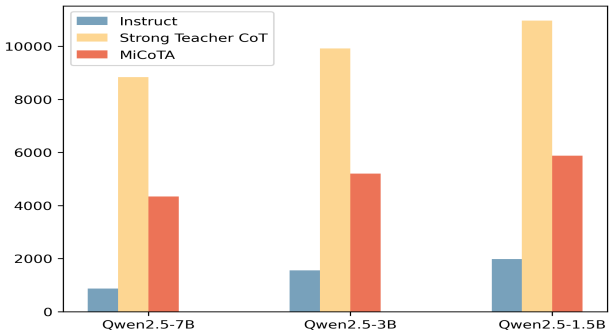

Figure 3: Comparison of Response Lengths (in tokens) for Different Methods across Student Models.

High Entropy Token	MiCoTA Rank	DeepSeek Rank
try	**22**	—
consider	**24**	—
by	**16**	—
given	**27**	—
from	9	**15**
wait	**13**	4
perhaps	—	**29**
actually	—	**30**
maybe	**17**	9
think	**23**	19

Table 6: High Entropy Token Ranking

5 Conclusion

In this paper, we propose Mid-CoT Teacher Assistant Distillation (MiCoTA), a novel approach designed to address the "SLMs Learnability Gap" problem by mitigating both the capacity gap between the student and teacher models, as well as the length gap by learning from intermediate-length CoT, allowing smaller models to better benefit from large reasoning LLMs. Our extensive experiments across multiple benchmarks demonstrate that MiCoTA significantly improves the reasoning performance of SLMs, achieving up to 35.6% better results compared to models directly trained on long CoT data from large teachers.

By providing thorough extensive experiments, MiCoTA paves the way for more effective training of smaller models while retaining strong reasoning capabilities. Our findings highlight the potential of intermediate-length CoT sequences in fostering better reasoning outcomes in resource-constrained environments, offering a promising direction for future research in sequence-level knowledge distillation and model efficiency.

References

Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Knowledge distillation: A good teacher is patient and consistent. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10925–10934, 2022.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Hugging Face. Open r1: A fully open reproduction of deepseek-r1, January 2025. URL https://github.com/huggingface/open-r1.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024. URL https://zenodo.org/records/12608602.

Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vladimir Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. Arcee's MergeKit: A toolkit for merging large language models. In Franck Dernoncourt, Daniel Preoţiuc-Pietro, and Anastasia Shimorina (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 477–485,

Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-industry.36. URL https://aclanthology.org/2024.emnlp-industry.36.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. Reasoning with language model is planning with world model. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 8154–8173. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.507. URL https://doi.org/10.18653/v1/2023.emnlp-main.507.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems, 2024. URL https://arxiv.org/abs/2402.14008.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset, 2021. URL https://arxiv.org/abs/2103.03874.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4163–4174. Association for Computational Linguistics, 2020.

Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 1317–1327, 2016.

Yuetai Li, Xiang Yue, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, Bhaskar Ramasubramanian, and Radha Poovendran. Small models struggle to learn from strong reasoners, 2025. URL https://arxiv.org/abs/2502.12143.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 5191–5198, 2020.

OpenAI. Learning to reason with llms, 2024a. URL https://openai.com/index/learning-to-reason-with-llms/.

OpenAI. Openai o1 system card, 2024b. URL https://cdn.openai.com/o1-system-card-20241205.pdf.

Geondo Park, Gyeongman Kim, and Eunho Yang. Distilling linguistic context for language model compression. In *Conference on Empirical Methods in Natural Language Processing*, 2021. URL https://api.semanticscholar.org/CorpusID:237563200.

Zhenting Qi, Mingyuan Ma, Jiahang Xu, Li Lyna Zhang, Fan Yang, and Mao Yang. Mutual reasoning makes smaller llms stronger problem-solvers. *CoRR*, abs/2408.06195, 2024. doi: 10.48550/ARXIV.2408.06195. URL https://doi.org/10.48550/arXiv.2408.06195.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2015.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

ByteDance Seed, Yufeng Yuan, Yu Yue, Mingxuan Wang, Xiaochen Zuo, Jiaze Chen, Lin Yan, Wenyuan Xu, Chi Zhang, Xin Liu, et al. Seed-thinking-v1. 5: Advancing superb reasoning models with reinforcement learning. *arXiv preprint arXiv:2504.13914*, 2025.

Wonchul Son, Jaemin Na, Junyong Choi, and Wonjun Hwang. Densely guided knowledge distillation using multiple teacher assistants. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9395–9404, 2021.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for bert model compression. *arXiv preprint arXiv:1908.09355*, 2019.

Raphael Tang, Yao Lu, Linqing Liu, Lili Mou, Olga Vechtomova, and Jimmy J. Lin. Distilling task-specific knowledge from bert into simple neural networks. *ArXiv*, abs/1903.12136, 2019. URL `https://api.semanticscholar.org/CorpusID:85543565`.

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

Open Thoughts Team. Open Thoughts, January 2025a.

Qwen Team. Qwen2.5: A party of foundation models, September 2024a. URL `https://qwenlm.github.io/blog/qwen2.5/`.

Qwen Team. Qwq: Reflect deeply on the boundaries of the unknown, November 2024b. URL `https://qwenlm.github.io/blog/qwq-32b-preview/`.

Qwen Team. Qwq-32b: Embracing the power of reinforcement learning, March 2025b. URL `https://qwenlm.github.io/blog/qwq-32b/`.

Ziyu Wan, Xidong Feng, Muning Wen, Stephen Marcus McAleer, Ying Wen, Weinan Zhang, and Jun Wang. Alphazero-like tree-search can guide large language model decoding and training. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL `https://openreview.net/forum?id=C4OpREezgj`.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning, 2025. URL `https://arxiv.org/abs/2506.01939`.

Tiannan Wang, Wangchunshu Zhou, Yan Zeng, and Xinsong Zhang. Efficientvlm: Fast and accurate vision-language models via knowledge distillation and modal-adaptive pruning. *arXiv preprint arXiv:2210.07795*, 2022.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788, 2020.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13484–13508, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.754. URL `https://aclanthology.org/2023.acl-long.754`.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

Liang Wen, Yunke Cai, Fenrui Xiao, Xin He, Qi An, Zhenyu Duan, Yimin Du, Junchen Liu, Lifu Tang, Xiaowei Lv, et al. Light-r1: Curriculum sft, dpo and rl for long cot from scratch and beyond. *arXiv preprint arXiv:2503.10460*, 2025.

Han Wu, Yuxuan Yao, Shuqi Liu, Zehua Liu, Xiaojin Fu, Xiongwei Han, Xing Li, Hui-Ling Zhen, Tao Zhong, and Mingxuan Yuan. Unlocking efficient long-to-short llm reasoning with model merging, 2025. URL https://arxiv.org/abs/2503.20641.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*, 2023.

Canwen Xu, Wangchunshu Zhou, Tao Ge, Furu Wei, and Ming Zhou. BERT-of-Theseus: Compressing BERT by progressive module replacing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7859–7869. Association for Computational Linguistics, 2020.

Canwen Xu, Wangchunshu Zhou, Tao Ge, Ke Xu, Julian J. McAuley, and Furu Wei. Beyond preserved accuracy: Evaluating loyalty and robustness of bert compression. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 10653–10659. Association for Computational Linguistics, 2021.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/271db9922b8d1f4dd7aaef84ed5ac703-Abstract-Conference.html.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. Limo: Less is more for reasoning, 2025. URL https://arxiv.org/abs/2502.03387.

Huifeng Yin, Yu Zhao, Minghao Wu, Xuanfan Ni, Bo Zeng, Hao Wang, Tianqi Shi, Liangying Shao, Chenyang Lyu, Longyue Wang, et al. Towards widening the distillation bottleneck for reasoning models. *arXiv preprint arXiv:2503.01461*, 2025.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch, 2024. URL https://arxiv.org/abs/2311.03099.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. Star: Bootstrapping reasoning with reasoning. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/639a9a172c044fbb64175b5fad42e9a5-Abstract-Conference.html.

Chen Zhang, Dawei Song, Zheyu Ye, and Yan Gao. Towards the law of capacity gap in distilling language models. *arXiv preprint arXiv:2311.07052*, 2023a.

Chen Zhang, Yang Yang, Jiahao Liu, Jingang Wang, Yunsen Xian, Benyou Wang, and Dawei Song. Lifting the curse of capacity gap in distilling language models. *CoRR*, abs/2305.12129, 2023b. doi: 10.48550/ARXIV.2305.12129. URL https://doi.org/10.48550/arXiv.2305.12129.

Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 11953–11962, 2022.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models, 2024. URL https://arxiv.org/abs/2403.13372.

Wangchunshu Zhou, Canwen Xu, and Julian McAuley. BERT learns to teach: Knowledge distillation with meta learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7037–7049. Association for Computational Linguistics, 2022.

Chenghao Zhu, Nuo Chen, Yufei Gao, Yunyi Zhang, Prayag Tiwari, and Benyou Wang. Is your llm outdated? a deep look at temporal generalization, 2025. URL https://arxiv.org/abs/2405.08460.

A Training Details

Table 7: List of all models used in our experiments, with Hugging Face links where available.

Model Name	Hugging Face Link
QwQ-32B (Team, 2025b)	https://huggingface.co/Qwen/QwQ-32B
R1-Distill-Qwen-32B (DeepSeek-AI et al., 2025)	https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-32B
Qwen2.5-32B-Instruct (Team, 2024a)	https://huggingface.co/Qwen/Qwen2.5-32B-Instruct
Qwen2.5-14B-Instruct	https://huggingface.co/Qwen/Qwen2.5-14B-Instruct
Qwen2.5-7B-Instruct	https://huggingface.co/Qwen/Qwen2.5-7B-Instruct
Qwen2.5-3B-Instruct	https://huggingface.co/Qwen/Qwen2.5-3B-Instruct
Qwen2.5-1.5B-Instruct	https://huggingface.co/Qwen/Qwen2.5-1.5B-Instruct

A.1 Models

Table 7 provides a detailed summary of the models utilized in our study.

A.2 Parameters Setting

Qwen2.5-14B-Instruct The model is trained using eight NVIDIA A800-SXM4-80GB GPUs. We set the batch size to 32 and the peak learning rate to 1e-5, following a cosine decay schedule. A weight decay of 0.01 is applied, and for all experiments, we train for a maximum of two epochs.

Qwen2.5-7B-Instruct The model is trained using eight NVIDIA A800-SXM4-80GB GPUs. For all experiments, we set the batch size to 32 and the peak learning rate to 1e-5, following a cosine decay schedule. A weight decay of 0.01 and a maximum of two epochs are applied.

Qwen2.5-3B-Instruct The model is trained using eight NVIDIA A800-SXM4-80GB GPUs. For the QwQ-32B strong teacher experiment, we set the batch size to 32 and the peak learning rate to 1e-5, following a cosine decay schedule. A weight decay of 0.01 and a maximum of two epochs are applied. For the R1-Distill-Qwen-32B strong teacher experiment, we set the batch size to 128 and the peak learning rate to 2e-5, also following a cosine decay schedule. A weight decay of 0.01 and a maximum of three epochs are applied.

Qwen2.5-1.5B-Instruct The model is trained using eight NVIDIA A800-SXM4-80GB GPUs. For the QwQ-32B strong teacher experiment, we set the batch size to 32 and the peak learning rate to 1e-5, following a cosine decay schedule. A weight decay of 0.01 and a maximum of two epochs are applied. For the R1-Distill-Qwen-32B strong teacher experiment, we set the batch size to 96 and the peak learning rate to 2e-5, also following a cosine decay schedule. A weight decay of 0.01 and a maximum of three epochs are applied.

B More Experiments Results

B.1 Generalization Across Architectures

We extend our experiments to models of the LLaMA family to demonstrate the generality of our approach to other architectures. Specifically, we use DeepSeek-R1-Distill-Llama-70B as the strong teacher, LLaMA-3.1-8B-Instruct as the teacher assistant, and LLaMA-3.2-3B-Instruct as the student model. As shown in Table 8, "Instruct" denotes the untrained checkpoint; "Strong teacher CoT" denotes the version trained on CoT data distilled from DeepSeek-R1-Distill-Llama-70B; "Merge" denotes the teacher assistant obtained by merging LLaMA-3.1-8B-Instruct with the checkpoint trained by the strong teacher CoT; and "MiCoTA " denotes the student model trained on data distilled from the Merge teacher assistant. As demonstrated in Table 8, the advantages of MiCoTA transfer effectively to the LLaMA family. The relative improvements show similar patterns to Qwen family, which further verifies the effectiveness of our proposed method.

Table 8: Ablation studies on MiCoTA with LLaMA-3.1-8B, LLaMA-3.2-3B. The highest score is bolded, and the second highest score is underlined.

Models	Method	AIME	AMC	Olympiad	MATH-500	GSM8K	Average
LLaMA-3.1-8B	Instruct	**3.33**	**35.00**	12.00	47.60	76.80	34.94
	Strong Teacher CoT	**3.33**	25.00	13.18	48.40	**81.12**	34.20
	MiCoTA	0	30.00	**14.66**	**55.60**	75.73	**35.19**
LLaMA-3.2-3B	Insturct	**3.33**	22.50	10.07	**46.40**	69.74	30.40
	Strong Teacher CoT	**3.33**	17.50	10.51	44.00	73.00	29.66
	MiCoTA	**3.33**	**27.50**	**10.81**	41.20	**76.04**	**31.77**

Table 9: Qwen-2.5-32B as Judge results; the highest score is bolded, and the second highest is underlined.

Models	Method	AIME	AMC	Olympiad	MATH-500	GSM8K	Average
Qwen2.5-7B	Instruct	10.00	40.00	**39.85**	**77.20**	**91.96**	51.80
	Strong Teacher CoT	6.66	27.50	29.48	67.40	90.14	44.24
	Mix-Long*	10.00	37.50	38.37	72.00	89.38	49.45
	MiCoTA	**13.33**	**52.50**	38.07	74.20	90.14	**53.65**
Qwen2.5-3B	Instruct	3.33	35.00	27.11	**65.60**	84.91	43.19
	Strong Teacher CoT	3.33	22.50	16.44	47.80	81.60	34.33
	Mix-Long*	10.00	37.50	28.74	63.40	82.71	44.47
	MiCoTA	**13.33**	**40.00**	**29.03**	64.20	**85.06**	**46.32**
Qwen2.5-1.5B	Instruct	**3.33**	22.50	**20.59**	46.80	**73.16**	33.28
	Strong Teacher CoT	0.00	7.50	10.81	34.20	65.65	23.63
	Mix-Long*	**3.33**	25.00	18.07	**54.4**	73.08	34.78
	MiCoTA	**3.33**	**27.50**	19.55	53.2	72.47	**35.21**

* The results presented here are based on our reproduction of the method.

B.2 LLM as Judge

Table 9 shows the detailed performance scores of each benchmark for different student models, with Qwen-2.5-32B serving as the judge. The results in the table indicate that our method effectively enhances their overall performance.

B.3 Merge Method Ablation

To further investigate the impact of merging strategies on performance, we conduct a merge method ablation study. We merge Qwen2.5-14B-Instruct with its Strong-teacher-CoT-trained counterpart using merge ratios of 0.5 and 0.5 (for the DARE method, we adopt 0.5 as target model density) to obtain 3 variants of the teacher assistant (TA), which are then used to distill the student model. As shown in Table 10, the performance is sensitive to the merging method selection. Specifically, the linear merging method attains the best performance for the 7B student model, with a score of 51.70. On the other hand, for the 3B and 1.5B student models, the DARE method outperforms both the linear and SLERP methods. Therefore, we adopt the DARE method in our paper to enhance the performance of small language models.

Table 10: Performance of Different Merge Methods across Model Sizes

Size	Linear	SLERP	DARE
Qwen2.5-7B	**51.70**	49.88	49.36
Qwen2.5-3B	41.25	40.17	**43.29**
Qwen2.5-1.5B	31.56	31.90	**33.01**

B.4 Effect of Merge Ratio

We adopt linear merging as the model merging strategy to explore the effect of merge ratio. We merge Qwen2.5-14B-Instruct with its Strong-teacher-CoT-trained counterpart using merge ratios of (0.9, 0.1), (0.1, 0.9) and (0.5,0.5) to obtain 3 variants of the teacher assistant (TA), which are then used to distill the student model.

As shown in Table 11, the experimental results reveal a clear pattern: a higher proportion of instruction (0.9Inst + 0.1CoT) yields better performance, especially for smaller student models. Specifically, for the 3B and 1.5B models, the (0.9Inst + 0.1CoT) configuration achieves the highest average scores (**44.08** and **33.56** respectively), outperforming both the balanced ratio (0.5:0.5) and the CoT-dominant ratio (0.1:0.9) by significant margins. Only the 7B model deviates from this trend, with the balanced ratio (0.5Inst + 0.5CoT) achieving the highest score (**52.70**), indicating that larger student models may have greater capacity to leverage both instruction and CoT knowledge synergistically.

Table 11: Average Score across Benchmarks (AIME2024, AMC, Olympiad, Math-500 and GSM8k)

Student Size	0.9Inst + 0.1CoT	0.5Inst + 0.5Cot	0.1Inst + 0.9 CoT
Qwen2.5-7B	48.10	**52.70**	44.19
Qwen2.5-3B	**44.08**	41.25	33.636
Qwen2.5-1.5B	**33.56**	31.56	29.038

C Case Study of Different Output

In this case study, we analyze the performance of the student model trained on Half-size CoT and MiCoTA, respectively. Half-size CoT is generated from Intermediate-sized Teacher Assistant and MiCoTA is generated from Mid-CoT Teacher Assistant. As shown in Figure 4, the models trained on MiCoTA and Half-size CoT both exhibit the thinking patterns characteristic of Strong Teacher CoT, particularly the strategy of reflection. However, the model trained on Half-size CoT exhibited excessive instances of "wait," amounting to 209 occurrences throughout its response, indicating a tendency to overanalyze along with becoming bogged down in repetitive checks. In contrast, the model trained on MiCoTA used "wait" only once. After providing the correct answer, it checked its solution briefly to ensure its accuracy, thereby concluding its response effectively. In summary, the model trained on MiCoTA not only reached the correct conclusion but also achieved a balanced CoT length.

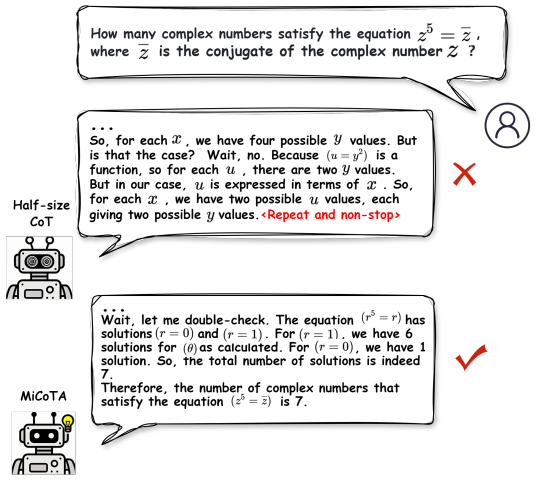

Figure 4: Case Study of MiCoTA. Models trained on Half-size CoT tend to overthink, whereas those trained on MiCoTA maintained a balanced reasoning path and reached the correct answer.