

Mentor-KD: Making Small Language Models Better Multi-step Reasoners

Hojae Lee^{1*}, Junho Kim^{2*}, SangKeun Lee^{1,2}

¹Department of Computer Science and Engineering ²Department of Artificial Intelligence
Korea University, Seoul, Republic of Korea
{221eehojae, monocrat, yalphy}@korea.ac.kr

Abstract

Large Language Models (LLMs) have displayed remarkable performances across various complex tasks by leveraging Chain-of-Thought (CoT) prompting. Recently, studies have proposed a Knowledge Distillation (KD) approach, *reasoning distillation*, which transfers such reasoning ability of LLMs through fine-tuning language models of multi-step rationales generated by LLM teachers. However, they have inadequately considered two challenges regarding insufficient distillation sets from the LLM teacher model, in terms of 1) data quality and 2) soft label provision. In this paper, we propose Mentor-KD, which effectively distills the multi-step reasoning capability of LLMs to smaller LMs while addressing the aforementioned challenges. Specifically, we exploit a mentor, intermediate-sized task-specific fine-tuned model, to augment additional CoT annotations and provide soft labels for the student model during reasoning distillation. We conduct extensive experiments and confirm Mentor-KD’s effectiveness across various models and complex reasoning tasks¹.

1 Introduction

Large Language Models (LLMs) have shown impressive emergent capabilities, showing their competence on a variety of reasoning tasks in the Natural Language Processing (NLP) landscape (Brown et al., 2020; Rae et al., 2021; Hoffmann et al., 2022; Chowdhery et al., 2023). One particularly interesting strategy for this approach is Chain-of-Thought (CoT) prompting, which elicits multi-step reasoning abilities of LLMs by explicitly generating intermediate reasoning steps for complex tasks (Wei et al., 2022b). However, such reasoning abilities have been shown to only manifest in language models (LMs) with over hundreds of billion param-

eters (Chung et al., 2022; Wei et al., 2022a), which require significant computational resources or expensive API calls, restricting their deployment on resource-limited scenarios.

To circumvent these deployment challenges, previous works (Ho et al., 2023; Li et al., 2023; Magister et al., 2023) have followed a knowledge distillation (KD) approach, *reasoning distillation*, which transfers the multi-step reasoning ability of LLMs to small LMs. The KD pipeline generally applies In-Context Learning (ICL) on the LLM teacher model to generate outputs (e.g., multi-step rationales) as distillation sets, and then utilizes them to fine-tune the student model. Previous studies have shown that reasoning distillation can significantly improve student performances and may even outperform their LLM teachers on specific tasks (Ho et al., 2023; Chen et al., 2023).

However, previous approaches to reasoning distillation have two challenges arising from insufficient distillation sets generated by LLM teachers. First, as LLMs may not have access to task-specific data, the quality of the rationales for distillation can be low (e.g., only 58% accuracy on GPT-3.5 for StrategyQA). The low quality of LLM teacher rationales limits the number of reasoning rationales to only a small set of correct ones due to the exclusion of incorrect rationales that negatively affect student performances (Ho et al., 2023). Second, because accessibility of black-box LLM teachers is generally restricted, the student model cannot mimic the predictive behavior and knowledge from the soft labels (Hinton et al., 2015). Such oversights may lead to the student model being over-fitted on limited distillation sets from teacher models and undermine its generalization capabilities.

To address these challenges, we propose Mentor-KD, a novel reasoning distillation framework that effectively distills the multi-step reasoning capability of LLMs. Our core idea is to introduce a mentor, an intermediate-sized task-specific model, that

* These authors contributed equally to this work.

¹Our code and data are available at <https://github.com/2hojae/mentor-kd>

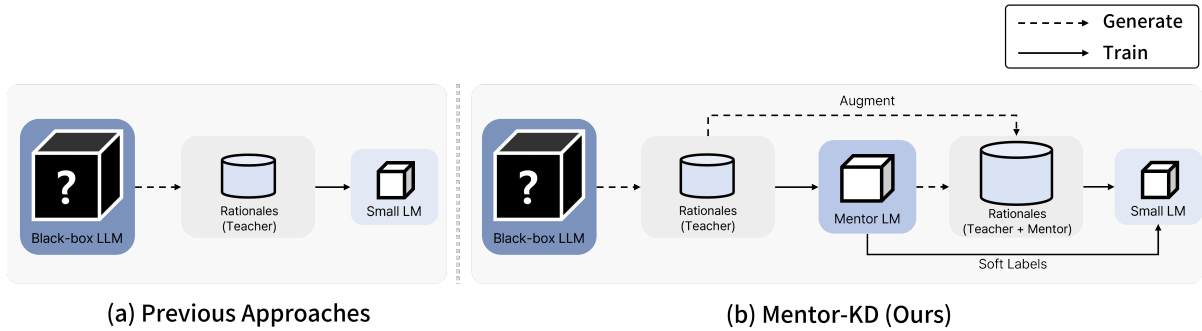


Figure 1: Comparison between (a) previous approaches of reasoning distillation and (b) Mentor-KD (ours). Our framework utilizes an intermediate-sized task-specific mentor model to complement the distillation sets of teachers.

complements the LLM teacher’s knowledge during reasoning distillation. To this end, we first fine-tune the mentor models on specific tasks and generate both CoT rationales and soft labels to augment distillation sets. By leveraging task-specific mentors whose power is concentrated toward a specific target ability, Mentor-KD effectively addresses two issues through training on more diverse rationales and intrinsic knowledge from soft labels.

We conduct extensive experiments on various types of complex reasoning tasks, including commonsense, arithmetic, logical, and symbolic reasoning tasks. The experimental results clearly demonstrate the superiority of our method over baselines leveraging knowledge only from LLMs. In addition, we verify that the mentor model can generate a substantial number of correct reasoning samples compared to other LLM baselines, highlighting the effectiveness of our method as means of data augmentation. Lastly, we demonstrate that our Mentor-KD significantly improves student performances in low-resource scenarios, indicating its cost-efficiency. In summary, the contributions of this paper include the following:

- We propose Mentor-KD, a novel reasoning distillation framework, which improves the reasoning ability of small LMs considering the limitations of insufficient distillation sets from LLM teachers.
- We introduce a mentor model to additionally generate both rationale samples and soft labels to complement the limited training datasets from the LLM teachers.
- We demonstrate that Mentor-KD improves the effectiveness of reasoning distillation on students with various types of reasoning and models through extensive experiments.

2 Related Works

2.1 Chain-of-Thought Prompting

CoT prompting is a method that elicits multi-step reasoning abilities of LMs through ICL (Wei et al., 2022b). The essence of CoT is that it acts as a guidance of logical progression for LMs to decompose and solve complex reasoning tasks (Xia et al., 2024). Consequently, it allowed LMs to excel in complex reasoning tasks (Kojima et al., 2022; Wang et al., 2023b; Zhang et al., 2023) which traditional few-shot learning methods have struggled with (Rae et al., 2021). Recent works take a step further to improve CoT prompting through enhancing the quality of reasoning steps. Madaan et al. (2023) had LMs to iteratively self-refine reasoning through self-feedback, while Gou et al. (2024) leveraged external tools for obtaining feedback. Trivedi et al. (2023); Zhao et al. (2023) incorporated information retrieval systems to enhance the facticity of LMs’ reasoning.

Despite the success, previous works (Hoffmann et al., 2022; Wei et al., 2022b; Chu et al., 2024) reported that the merits of reasoning on CoT prompting emerge when LMs are scaled to hundreds of billions of parameters. To address such problems, our work focuses on enabling CoT reasoning to small-scaled LMs through reasoning distillation.

2.2 Knowledge Distillation for LLMs

KD (Hinton et al., 2015) has been proven to be a promising approach to compress LMs by transferring the predictive behavior (e.g., soft labels) or internal knowledge (e.g., hidden representations) from larger LMs to smaller ones. However, existing KD methods for pre-trained LMs, which involve distilling the soft labels (Sanh et al., 2019; Gu et al., 2024) or representations (Wang et al., 2020, 2021;

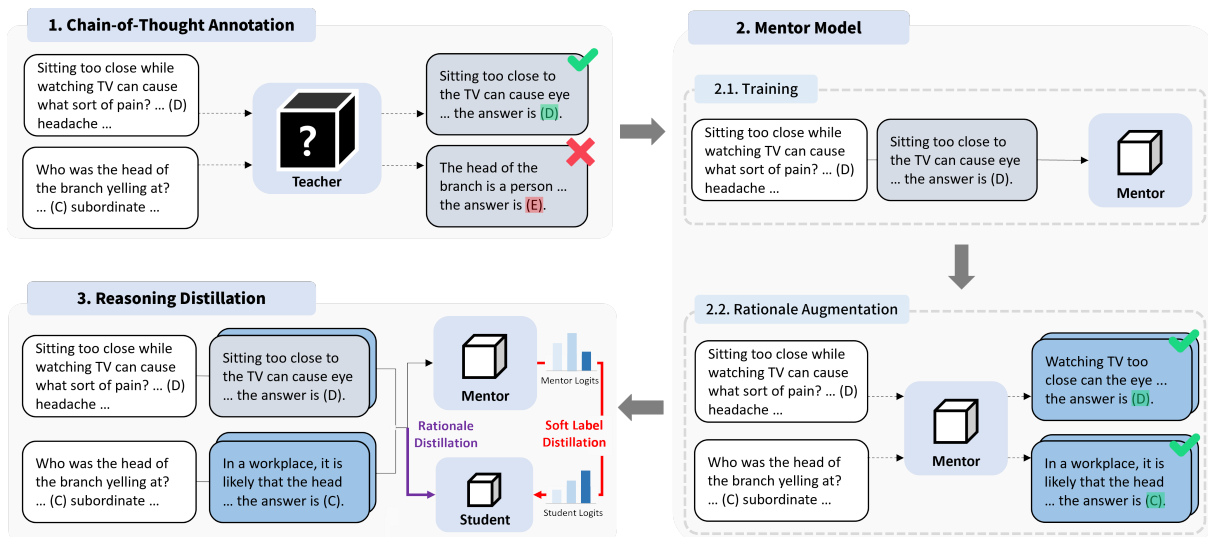


Figure 2: A general overview of our proposed framework, Mentor-KD. Mentor-KD is composed of three steps. First, CoT annotations are initially collected from the teacher LLM and filtered. Second, the preserved annotations are used to train the mentor model, and the trained mentor model augments multi-step rationales. Lastly, the student model is trained on annotations from the teacher and the student, as well as soft labels from the mentor model.

Kim et al., 2022), require access to the internal parameters of teachers. These requirements pose a significant challenge for leveraging LLMs in KD, regarding their black-box nature and impracticality.

In turn, recent works practiced reasoning distillation, which enabled smaller LMs (students) to carry out multi-step reasoning similar to LLMs by utilizing rationales generated by LLM teachers instead of soft labels. For example, Ho et al. (2023); Magister et al. (2023); Li et al. (2023) fine-tuned students on multi-step rationales that LLMs generated. Similarly, Shridhar et al. (2023) had students learn how to decompose a complex question through having LLMs to generate sub-problems to the original question. Wang et al. (2023c) iteratively employed LLMs to provide real-time feedback specifically tailored to the student’s generations. Kang et al. (2023); Zhao et al. (2024) leveraged information retrieval systems to enhance the facticity of student’s reasoning on knowledge-intensive tasks. Recently, Zhu et al. (2024a,b) incorporated multi-step rationales in a code format generated from the LLMs to improve the student’s arithmetic reasoning skills. Contemporaneous to our work, (Zhou and Ai, 2024) also utilized intermediate-sized models for LLM distillation. Our work differs in that we use intermediate-sized models for complementing the teacher model’s distillation signals, rather than for filtering the annotations.

While most previous works have been conducted to improve reasoning distillation by utilizing distil-

lation sets provided by LLMs, we posit that they may be insufficient and may undermine the student’s capabilities. In this sense, our work is different in that we complement such insufficiency of LLM teachers.

3 Methodology

We elaborate on the detailed implementations of our Mentor-KD. The core idea is to augment the distillation training set by leveraging a task-specific intermediate-sized mentor model. To this end, we first generate CoT annotations from LLM teacher models (Section 3.1). We then fine-tune the mentor model with the distillation set from the LLM teacher, and the trained mentor model generates additional training sets, including both rationales and soft labels (Section 3.2). By augmenting both signals from the mentor, we distill the knowledge to student models (Section 3.3). Figure 2 illustrates an overview of our framework.

3.1 Chain-of-Thought Annotations

We use the LLM to obtain CoT annotations composed of a rationale and a final prediction to a question via Zero-shot-CoT (Kojima et al., 2022). It is a two-staged strategy consisting of reasoning and answer extraction stages, and thus, we induce the LLM to generate a CoT rationale first and subsequently a final prediction afterwards.

Specifically, we first append “Let’s think step by step” to the question and prompt the LLM to obtain

the rationale. In sequence, we prompt the LLM again by incorporating the previously obtained rationale to induce its final prediction. Formally, from a dataset $\mathcal{D} = \{q_i, y_i\}$ where q_i denotes a question and y_i denotes a golden label, our goal is to induce the LLM to generate a step-by-step rationale r_i^t and a final prediction \hat{y}_i^t , given q_i as an input. The prompting template takes the form of: "Q: $\{q_i\}$. A: Let's think step by step. $\{r_i^t\}$. Therefore, the answer is $\{\hat{y}_i^t\}$ ".

Afterward, we filter the annotations generated by the LLM. Following previous works (Li et al., 2023; Magister et al., 2023; Fu et al., 2023; Lee et al., 2024), we preserve annotations where the final prediction \hat{y}_i^t matches the golden answer y_i of a sample. Then, the annotations are reformatted into a question-label format for training mentor and student models. More formally, for all annotations i where $\hat{y}_i^t = y_i$, we reformat a data sample $(q_i, r_i^t, \hat{y}_i^t, y_i)$ into (q_i, l_i^t, y_i) , where l_i^t takes the form of " $\{r_i^t\}$ ". Therefore, the answer is $\{y_i\}$." Consequently, we finally construct $\mathcal{D}_{\text{teacher}} = \{(q_i, l_i^t, y_i)\}_{i=1}^N$.

3.2 Mentor Model

Here, we describe how our mentor models are trained to concentrate their powers to a specific task, and utilized to complement the insufficient distillation sets of LLM teachers.

Training. For training the mentor model, we directly fine-tune it on the previously constructed $\mathcal{D}_{\text{teacher}}$. Specifically, the mentor model receives q_i as an input, l_i^t as a label, and is trained with a standard language modeling objective.

Rationale Augmentation. The trained mentor model is then used for train data augmentation. For data samples from \mathcal{D} , we let the mentor model annotate step-by-step rationales, given q_i as an input. The mentor in return generates a label l_i^m , which consists of a step-by-step rationale and a prediction of its own. We filter the annotations by the mentor identical to filtering the teacher's annotations and preserve data samples where $\hat{y}_i^m = y_i$. Through this stage, we construct $\mathcal{D}_{\text{mentor}} = \{(q_i, l_i^m, y_i)\}_{i=1}^N$ per dataset.

With annotations obtained from the teacher ($\mathcal{D}_{\text{teacher}}$) and the mentor ($\mathcal{D}_{\text{mentor}}$), we finally construct $\mathcal{D}_{\text{train}}$ for training the student model², which

²It is worth noting that we do not distinguish where the CoT annotations were generated from, but we randomly sample instances from $\mathcal{D}_{\text{train}}$ to train the student models.

is defined as follows:

$$\mathcal{D}_{\text{train}} = \mathcal{D}_{\text{teacher}} \cup \mathcal{D}_{\text{mentor}} \quad (1)$$

3.3 Reasoning Distillation

For training the student model, we incorporate both fine-tuning (rationale distillation) and knowledge distillation through logit values obtainable via the mentor model (soft label distillation). This is to allow the student model to jointly 1) learn how to practice step-by-step reasoning in a symbolic manner (Ho et al., 2023; Li et al., 2023; Magister et al., 2023), as well as 2) mimic the predictive behavior of a larger model (Hinton et al., 2015). In correspondence, our training objective consists of two loss functions.

Rationale Distillation. Identical to training the mentor model, the step-by-step reasoning ability can be distilled through fine-tuning the student model with question-label pairs obtained from the teacher and the mentor. More specifically, the form of learning the multi-step reasoning ability through fine-tuning is defined as follows:

$$\mathcal{L}_{\text{rd}} = \mathbb{E}_{\mathcal{D}_{\text{train}}} \log P_f([q; r; y]), \quad (2)$$

where f indicates the student model, and the square brackets indicate string concatenation.

Soft Label Distillation. Leveraging the LLM teacher's internal knowledge can be impractical due to its black-box nature or enormous size. Instead, we employ our mentor model to provide the soft labels for distillation. The soft labels are obtained through a forward pass, followed by a softmax function, given q as an input. Formally, we obtain the soft label (probability distribution) p_k of the mentor and student models from the logit value z_k at the k -th position through the following equation:

$$p_k = \frac{\exp(z_k/\tau)}{\sum_j \exp(z_j/\tau)}, \quad (3)$$

where τ indicates a temperature hyperparameter for softening the distribution. After obtaining probability distributions of the mentor (p^m) and the student (p^s), we adopt the Kullback-Leibler divergence loss to minimize the divergence between the two distributions. This allows the student model to mimic the predictive behavior and learn the internal

Model	#Params	GSM8K	ASDiv	SVAMP	CommonsenseQA
GPT-3.5-Turbo (teacher)*	-	73.98	79.64	75.14	74.35
FlanT5-XXL (mentor)	11B	34.34	50.32	51.71	85.01
GPT-3-curie (Ho et al., 2023)	6.7B	6.75	-	12.67	56.76
T5-XXL (Magister et al., 2023)	11B	21.99	42.12	-	-
FlanT5-XL (Fu et al., 2023)	3B	22.40	28.40	23.80	-
FlanT5-XL (Vanilla-KD)*	3B	22.76	29.41	29.33	81.13
FlanT5-XL (MCC-KD)*	3B	24.28	31.35	30.00	82.88
FlanT5-XL (Mentor-KD (ours))	3B	24.76	31.86	32.70	87.14

Table 1: Comparison with different baselines on arithmetic and commonsense reasoning tasks. The reported results are averaged accuracy over four runs using randomly selected seeds. Performances marked with an asterisk(*) were excerpted from MCC-KD (Chen et al., 2023). The best results are highlighted in **boldface**.

knowledge of larger models. The training objective for soft label distillation is defined as follows:

$$\mathcal{L}_{\text{sld}}(p^m, p^s) = \sum_k p_k^m \log \frac{p_k^m}{p_k^s} \quad (4)$$

Joint Learning. Finally, we have the student model to jointly learn the aforementioned two objectives. The loss function for training the student model is as follows:

$$\mathcal{L} = (1 - \lambda)\mathcal{L}_{\text{rd}} + \lambda\mathcal{L}_{\text{sld}}, \quad (5)$$

where λ is a hyperparameter for interpolating the two loss functions.

4 Experiments

In this section, we describe the experiment details and evaluate our Mentor-KD on various complex reasoning tasks.

4.1 Experiment Setup

Tasks and Datasets. Following (Wei et al., 2022b; Kojima et al., 2022), we evaluate our Mentor-KD on four categories of complex reasoning tasks, which are commonsense, arithmetic, logical, and symbolic reasoning. Specifically, we adopt up to three datasets per task in order to evaluate our framework on various datasets of the same task type. Datasets used for this paper are StrategyQA (Geva et al., 2021), CommonsenseQA (Talmor et al., 2019) for commonsense reasoning, GSM8K (Cobbe et al., 2021), ASDiv (Miao et al., 2020), and SVAMP (Patel et al., 2021) for arithmetic reasoning, Tracking Shuffled Objects, Date Understanding (Srivastava et al., 2023) for logical reasoning, and Last Letter Concatenation (Wei et al., 2022b; Kojima et al., 2022) for symbolic reasoning. Further details are provided in Appendix A.

Language Models. We utilize gpt-3.5-turbo through OpenAI API for our teacher model. For the mentor and student models, we mainly use FlanT5-XXL and FlanT5-XL (Chung et al., 2022) as our mentor and student models. For additional analysis, we use various sizes of FlanT5 and T5 (Raffel et al., 2020), including large, base, and small-sized models.

Chain-of-Thought Annotations. For GSM8K, ASDiv, SVAMP, and CommonsenseQA, we utilize the CoT annotations provided by (Chen et al., 2023). The annotations were collected with GPT-3.5-Turbo using Zero-shot-CoT prompting, which is identical to our methodology mentioned in Section 3.1. Other datasets were newly prompted and collected by our research institute.

Baselines. For the baselines, we incorporate previous methods of reasoning distillation. Specifically, we implement Vanilla-KD, a general reasoning distillation method that fine-tunes student models on the teacher model’s generated rationales (Ho et al., 2023; Magister et al., 2023), and MCC-KD, which further emphasizes diversity and consistency within multiple CoT rationales (Chen et al., 2023). We also compare Mentor-KD’s performances with Fu et al. (2023), which aims to specialize LM’s reasoning ability towards a specific task. We report the teacher model’s performances via Zero-shot-CoT (ZS-CoT) prompting.

Implementations. We adopt models provided by HuggingFace (Wolf et al., 2020) on two NVIDIA RTX A6000 GPUs. Specifically, we train models for 18 epochs for XXL-/XL-sized models, 10 epochs for large, and 20 epochs for base, and small

Model	#Params	Method	Commonsense		Arithmetic		Logical		Symbolic
			SQA	CSQA	ASDiv	SVAMP	Shuffled	Date	Last Letter
GPT-3.5-Turbo	-	ZS-CoT (teacher)	58.07	74.35*	79.64*	75.14*	64.00	81.98	68.00
T5-large	780M	Vanilla-KD (mentor)	63.32	68.80	12.42	13.05	90.22	84.68	68.00
		Vanilla-KD	61.43	55.53	<u>11.15</u>	10.00	<u>77.33</u>	89.19	56.00
T5-base	250M	MCC-KD	<u>62.01</u>	<u>57.17</u>	9.55	8.00	56.89	81.98	45.33
		Mentor-KD (ours)	62.45	59.05	12.10	10.00	92.00	<u>88.29</u>	65.33
		Vanilla-KD	55.60	<u>42.75</u>	5.10	6.67	<u>39.11</u>	<u>81.98</u>	<u>48.67</u>
T5-small	80M	MCC-KD	<u>56.77</u>	<u>38.25</u>	<u>5.73</u>	<u>7.33</u>	38.22	77.48	28.67
		Mentor-KD (ours)	57.93	45.37	7.01	8.67	79.56	87.39	56.67
		Vanilla-KD	55.60	<u>42.75</u>	5.10	6.67	<u>39.11</u>	<u>81.98</u>	<u>48.67</u>

Table 2: Performances of teacher, mentor, and student models across four different complex reasoning tasks, where the backbone model is T5. GPT-3.5-Turbo results with an asterisk(*) were excerpted from (Chen et al., 2023). The best and second best results are highlighted in **boldface** and underline, respectively.

Model	#Params	Method	Commonsense		Arithmetic		Logical		Symbolic
			SQA	CSQA	ASDiv	SVAMP	Shuffled	Date	Last Letter
GPT-3.5-Turbo	-	ZS-CoT (teacher)	58.07	74.35*	79.64*	75.14*	64.00	81.98	68.00
FlanT5-large	780M	Vanilla-KD (mentor)	64.48	79.36	20.70	14.00	90.22	88.29	65.33
		Vanilla-KD	62.74	62.33	12.42	10.67	<u>84.89</u>	<u>86.49</u>	<u>53.33</u>
FlanT5-base	250M	MCC-KD	<u>64.92</u>	68.47	<u>13.69</u>	12.00	69.78	85.59	46.00
		Mentor-KD (ours)	65.21	<u>67.24</u>	15.29	<u>11.33</u>	93.78	87.39	65.33
		Vanilla-KD	55.90	<u>48.24</u>	<u>7.96</u>	10.67	<u>63.11</u>	85.59	<u>52.67</u>
FlanT5-small	80M	MCC-KD	<u>58.37</u>	45.21	7.01	<u>10.00</u>	43.11	81.98	35.33
		Mentor-KD (ours)	59.97	48.98	10.83	10.67	82.67	<u>83.78</u>	58.67
		Vanilla-KD	55.90	<u>48.24</u>	<u>7.96</u>	10.67	<u>63.11</u>	85.59	<u>52.67</u>

Table 3: Performances of teacher, mentor, and student models across four different complex reasoning tasks, where the backbone model is FlanT5. GPT-3.5-Turbo results with an asterisk(*) were excerpted from (Chen et al., 2023). The best and second best results are highlighted in **boldface** and underline, respectively.

Model	Method	Shuffled	Last Letter
T5	Mentor-KD (ours)	79.56	56.67
	w/o RD	32.89	50.00
	w/o SLD	76.00	52.00
FlanT5	Mentor-KD (ours)	82.67	58.67
	w/o RD	64.89	56.00
	w/o SLD	82.22	54.00

Table 4: Ablation study of Mentor-KD on Tracking Shuffled Objects and Last Letter Concatenation. We employ large models of each backbone model as mentors and small models as students.

models following the previous works (Chen et al., 2023; Ho et al., 2023). The maximum sequence length is set to 512 throughout all our experiments, and we sweep batch sizes in {2, 4, 6, 8}. To accelerate training and conserve memory usage, we apply mixed precision of bfloat16 and LoRA (Hu et al., 2022) throughout our main experiments and follow the related configurations from (Chen et al., 2023). Moreover, We use AdamW (Loshchilov and Hutter, 2019) optimizer, with a learning rate of {1e-4, 2e-4, 3e-4, 5e-4}. We apply the loss interpolation hyperparameter λ to 0.3, and the distillation temperature τ to {1.0, 2.0}. We report the average test accuracy results from four random seeds.

4.2 Main Results

For a fair comparison, we mainly compare Mentor-KD utilizing FlanT5-XL models on three arithmetic reasoning tasks and one commonsense reasoning task, which are commonly used in reasoning distillation (Ho et al., 2023; Chen et al., 2023). The main results are provided in Table 1. We observe that our Mentor-KD achieves state-of-the-art performance on four different reasoning datasets. Specifically, our model achieves approximately 2.0% better performance on averaged accuracy than MCC-KD, the previous SOTA model. The results demonstrate the effectiveness of Mentor-KD in addressing challenging complex reasoning tasks, including both arithmetic and commonsense reasoning.

5 Analysis

To delve into the benefits of our method, we perform a series of fine-grained analytical experiments with the following research questions (RQs):

- **RQ1.** Can Mentor-KD be generalized to the various sizes and types of student models? (§5.1)
- **RQ2.** How does each component in Mentor-KD contribute to its overall performance? (§5.2)

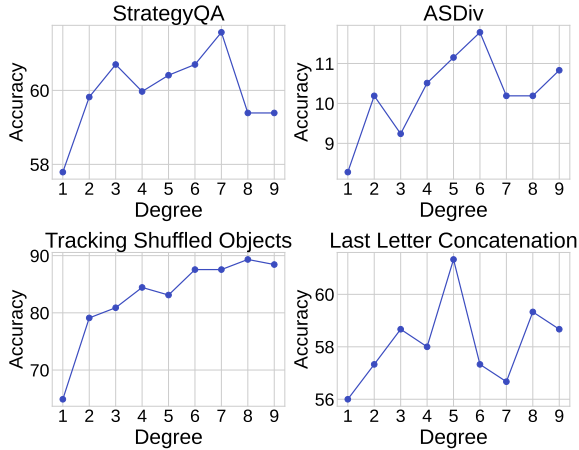


Figure 3: Performances by differentiating the degree (number) of mentor-generated CoT rationales per question. We adopt FlanT5-large and FlanT5-small as mentor and student models, respectively.

- **RQ3.** Can the mentor model generate informative distillation sets for students? (§5.3)
- **RQ4.** Does Mentor-KD offer improvements under low-resource scenarios? (§5.4)
- **RQ5.** Does the size of mentor models affect the performance of student models? (§5.5)

5.1 Various Student Models (RQ1)

To further investigate the generality of our Mentor-KD, we conduct experiments on various types of student models with different sizes. Notably, we further expand our scope of experiments by additionally incorporating logical and symbolic reasoning tasks. Specifically, we utilize T5 and FlanT5, which are widely adopted in LLM distillation following previous works (Ho et al., 2023; Chen et al., 2023). We leverage large variants of T5 and FlanT5 as our mentor model, and {base, small} variants as our student model. Details on implementations of this section are elaborated in Appendix B.

The results are shown in Tables 2 and 3. We observe that our Mentor-KD consistently outperforms the other baselines in four categories of complex reasoning tasks on various student models. In particular, Mentor-KD has shown large performance improvements in commonsense and logical reasoning tasks, which the student model may even outperform the performances of the LLM teacher (i.e., GPT-3.5). These results demonstrate that our task-specific mentor model can successfully complement the insufficient LLM teacher’s knowledge, thereby leading to achieving better performances for various student models by transferring more

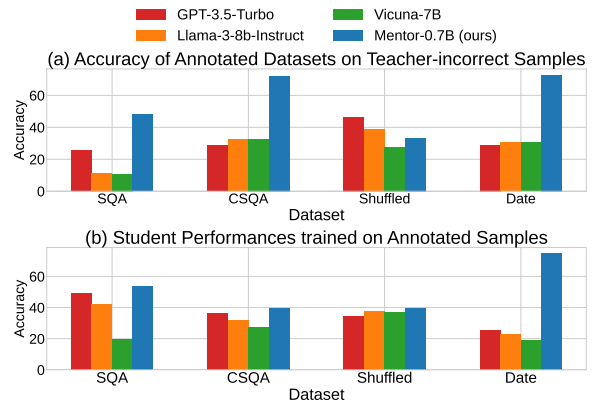


Figure 4: Comparison of (a) accuracy of our mentor model (FlanT5-large) and LLM baselines on teacher-incorrect samples, and (b) performances of student models trained with augmented distillation sets from LLM baselines and our mentor models.

informative distillation signals.

5.2 Ablation Studies (RQ2)

We conduct ablation studies to explore the contributions brought by each technique of our method. Specifically, we focus on the effect of rationale distillation (RD) and soft label distillation (SLD) from the mentor model. The detailed results are shown in Table 4. We observe that omitting RD and SLD significantly affects both model types and datasets. These results emphasize the significance of RD for both training samples and soft labels, which enhance the insufficient knowledge from teachers.

5.3 Impact of Data Augmentation (RQ3)

To further investigate the proposed data augmentation methods of mentor models, we additionally analyze the effectiveness in perspectives of both quantity and quality.

Quantity of Augmented Dataset. We first analyze the impact of the number of generated distillation sets from the mentor by diversifying the number of rationales that the mentor produces per question. The results are shown in Figure 3. Generally, we observe that student performances improve in line with the quantity of distillation sets. This indicates that our mentor models successfully generate rationales helpful for student models to learn multi-step reasoning. However, we also observe the performance usually saturated over six augmentations and begins to decline when more distillation sets are introduced, which may be due to the noises generated from models (Liu et al., 2022).

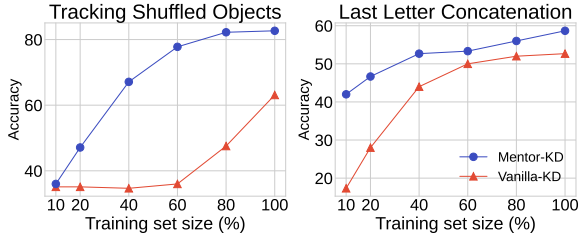


Figure 5: Comparison between Mentor-KD (Ours) and Vanilla-KD baseline on various distillation sets by differentiating the percentage of rationales being used.

Quality of Augmented Dataset. To investigate the quality of our augmented distillation sets, we compare our mentor models (i.e., FlanT5-large) with various LLMs that may be potential alternatives of mentors for augmentation (i.e. GPT-3.5-Turbo³, Llama-3-8B-Instruct⁴, and Vicuna-7B (Chiang et al., 2023)). We first compare the accuracy of the augmentations mentors generate with other baselines (through Zero-shot-CoT prompting) on incorrect samples predicted by the LLM teacher. We then report the performances of the student (i.e., FlanT5-small) trained on each augmentation to analyze whether task-specific mentors can provide informative sets to the students.

The results are shown in Figure 4. While the mentor models consist of smaller parameters than the LLMs (e.g., 10× smaller than Llama3-8B-Instruct), they generate more accurate rationales than other LLM baselines, indicating the ability to provide more diverse rationales for student models. In addition, we observe that the students trained with distillation sets from mentor models indeed achieve higher performance than those trained with sets from LLM teachers. These results suggest that mentors can generate higher-quality rationales than LLM teachers. Overall results highlight the superiority of task-specific fine-tuning of mentor models.

5.4 Low-resource Scenarios (RQ4)

In reasoning distillation, collecting sufficiently large distillation sets can be prohibitively expensive due to the cost of API calls for black-box LLMs. Therefore, we examine the effectiveness of Mentor-KD on low-resource scenarios, where distillation sets are collected for only a proportion of the original datasets. Specifically, we compare the Vanilla-

³We adopt a different seed value from the initial CoT annotation phase (Section 3.1) for this experiment.

⁴<https://ai.meta.com/blog/meta-llama-3/>

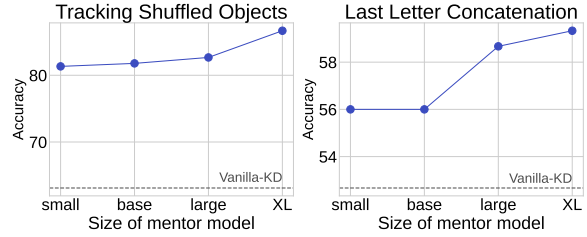


Figure 6: Comparison between student (FlanT5-small) performance using different mentor models considering various capacity gap sizes. Dotted lines in gray indicate Vanilla-KD baseline performances.

KD baseline with our Mentor-KD, varying the ratio of distillation sets generated from LLM teacher models. The results are shown in Figure 5.

We observe that the Mentor-KD also allows performance improvements for student models in low-resource scenarios, given that mentor models provide informative rationale sets and soft labels. In particular, the Vanilla-KD baseline shows performance degradation on highly limited distillation signals, while our Mentor-KD exhibits robustness for limited datasets. These results demonstrate that our mentor models can alleviate over-fitting problems for students from the limited distillation signals and can distill the LLM teacher’s knowledge in a cost-efficient manner. We elaborate on this research question in Appendix C.

5.5 Effects of Mentor Sizes (RQ5)

To further explore Mentor-KD’s effectiveness and verify our design choice, we conduct an additional experiment by differentiating the size of mentor models. Here, we employ FlanT5-small as a student model and FlanT5-{XL, large, base, small} as mentor models. For distilling small to small models, we utilize self-distillation, following previous works (Allen-Zhu and Li, 2023; Zhu et al., 2024a).

Figure 6 displays the results. Generally, we observe that the student model performs better when larger mentor models are incorporated during reasoning distillation. Employing the smallest mentor results in a performance decline, but we observe such scenarios still outperform the baselines in Table 3. The results suggest that employing larger models of better performances contributes to boosting the small student models’ performances, which is aligned with previous findings that student performances are correlated to their corresponding model’s performances (Ho et al., 2023).

6 Conclusion

We have presented Mentor-KD, a novel framework to transfer reasoning capabilities from LLMs to smaller LMs. To this end, we have introduced a mentor model, a novel auxiliary model, for complementing the distillation sets from LLMs by augmenting multi-step rationales and providing soft labels for the student model. Through extensive experiments, we have demonstrated that our Mentor-KD significantly improves the effectiveness of reasoning distillation. Specifically, our student models outperform existing reasoning distillation baselines with various sizes and types of models on complex reasoning tasks. Furthermore, we have verified that our mentor model can generate effective reasoning samples and soft labels for training student models, resulting in consistent performance improvements.

7 Limitations

While we have demonstrated that Mentor-KD effectively improves the reasoning ability of small language models by augmenting both training sets and soft labels, there are some limitations that present promising avenues for future research.

Training Costs for Mentor Models. Our framework requires additional computational costs for training mentor models for reasoning distillation. Besides the training costs in the distillation process, this study mainly focuses on improving the inference efficiency of small student models, as with most reasoning distillation research (Ho et al., 2023; Chen et al., 2023; Wang et al., 2023a). We further elaborate on this issue in Appendix C.

Exploration on Different Reasoning Strategies. While we successfully demonstrate the performance improvements in CoT reasoning abilities for small language models, it is an open question whether our framework can be applied to other types of reasoning strategies, such as program-guided reasoning (Zhu et al., 2024a), retrieval-based reasoning (Kang et al., 2023; Zhao et al., 2024), and reasoning based on contextualized, structured knowledge (Park et al., 2024). We leave the exploration of distillation for various types of reasoning strategies as a future research direction in this field.

Exploration on Different Architectures. We have verified the effectiveness of our framework on encoder-decoder models (e.g., FlanT5, T5) with

fewer than 3 billion parameters as the student models. Therefore, the applicability of our framework to decoder-only models remains under-explored in this work. Nevertheless, based on recent evidence suggesting that reasoning distillation can be effectively generalized to various architectures (Ho et al., 2023; Chen et al., 2023; Wang et al., 2023c), we believe that Mentor-KD is expected to display performance boosts on decoder-based student models as well.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.RS-2024-00415812 and No.2021R1A2C3010430) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2024-00439328, Karma: Towards Knowledge Augmentation for Complex Reasoning (SW Starlab), No.RS-2024-00457882, AI Research Hub Project, and No.RS-2019-II190079, Artificial Intelligence Graduate School Program (Korea University)).

References

- Zeyuan Allen-Zhu and Yuanzhi Li. 2023. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*. OpenReview.net.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Matiej Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Hongzhan Chen, Siyue Wu, Xiaojun Quan, Rui Wang, Ming Yan, and Ji Zhang. 2023. MCC-KD: Multi-CoT consistent knowledge distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6805–6820. Association for Computational Linguistics.
- 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. 2023. [Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality](#).
- 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 Sankaranarayanan 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. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24:240:1–240:113.
- Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2024. Navigate through enigmatic labyrinth a survey of chain of thought reasoning: Advances, frontiers and future. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1173–1203. Association for Computational Linguistics.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*.
- 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. 2021. Training verifiers to solve math word problems. *CoRR*.
- Dujian Ding, Ankur Mallick, Chi Wang, Robert Sim, Subhabrata Mukherjee, Victor Rühle, Laks V. S. Lakshmanan, and Ahmed Hassan Awadallah. 2024. Hybrid LLM: Cost-efficient and quality-aware query routing. In *ICLR*. OpenReview.net.
- Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023. Specializing smaller language models towards multi-step reasoning. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 202, pages 10421–10430. PMLR.
- Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.
- Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yuju Yang, Nan Duan, and Weizhu Chen. 2024. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *Proceedings of the International Conference on Learning Representations (ICLR)*. OpenReview.net.
- Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. Minillm: Knowledge distillation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*. OpenReview.net.
- Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *CoRR*.
- Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. Large language models are reasoning teachers. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 14852–14882. Association for Computational Linguistics.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack William Rae, and Laurent Sifre. 2022. An empirical analysis of compute-optimal large language model training. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*. OpenReview.net.
- Minki Kang, Seanie Lee, Jinheon Baek, Kenji Kawaguchi, and Sung Ju Hwang. 2023. Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc.
- Junho Kim, Jun-Hyung Park, Mingyu Lee, Wing-Lam Mok, Joon-Young Choi, and SangKeun Lee. 2022. Tutoring helps students learn better: Improving knowledge distillation for BERT with tutor network. In *Proceedings of the Conference on Empirical*

- Methods in Natural Language Processing (EMNLP)*, pages 7371–7382. Association for Computational Linguistics.
- Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 22199–22213. Curran Associates, Inc.
- Jooyoung Lee, Fan Yang, Thanh Tran, Qian Hu, Emre Barut, and Kai-Wei Chang. 2024. Can small language models help large language models reason better?: Lm-guided chain-of-thought. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2835–2843. ELRA and ICCL.
- 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. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2665–2679. Association for Computational Linguistics.
- Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022. Generated knowledge prompting for commonsense reasoning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3154–3169. Association for Computational Linguistics.
- Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*. OpenReview.net.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhunoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.
- Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2023. [Teaching small language models to reason](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1773–1781, Toronto, Canada. Association for Computational Linguistics.
- Shen-yun Miao, Chao-Chun Liang, and Keh-Yih Su. 2020. A diverse corpus for evaluating and developing English math word problem solvers. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 975–984. Association for Computational Linguistics.
- Jun-Hyung Park, Mingyu Lee, Junho Kim, and SangKeun Lee. 2024. Coconut: Contextualized commonsense unified transformers for graph-based commonsense augmentation of language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5815–5830. Association for Computational Linguistics.
- Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP models really able to solve simple math word problems? In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 2080–2094. Association for Computational Linguistics.
- Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Mari-beth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsim-poukelli, Nikolai Grigorev, Doug Fritz, Thibault Sotiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d’Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew J. Johnson, Blake A. Hechtman, Laura Weidinger, Jason Gabriel, William Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorraine Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling language models: Methods, analysis & insights from training gopher. *CoRR*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*.
- 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.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch,

- Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, and et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4149–4158. Association for Computational Linguistics.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037, Toronto, Canada. Association for Computational Linguistics.
- Zhongwei Wan, Xin Wang, Che Liu, Samiul Alam, Yu Zheng, Jiachen Liu, Zhongnan Qu, Shen Yan, Yi Zhu, Quanlu Zhang, Mosharaf Chowdhury, and Mi Zhang. 2024. Efficient large language models: A survey. *Transactions on Machine Learning Research*.
- Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. 2023a. **SCOTT: Self-consistent chain-of-thought distillation**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5546–5558, Toronto, Canada. Association for Computational Linguistics.
- Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2021. Minilmv2: Multi-head self-attention relation distillation for compressing pre-trained transformers. In *Findings of the Association for Computational Linguistics: ACL 2021*, pages 2140–2151. Association for Computational Linguistics.
- Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*. OpenReview.net.
- Zhaoyang Wang, Shaohan Huang, Yuxuan Liu, Jiahai Wang, Minghui Song, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2023c. Democratizing reasoning ability: Tailored learning from large language model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1948–1966. Association for Computational Linguistics.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 24824–24837. Curran Associates, Inc.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP)*, pages 38–45. Association for Computational Linguistics.
- Yu Xia, Rui Wang, Xu Liu, Mingyan Li, Tong Yu, Xiang Chen, Julian J. McAuley, and Shuai Li. 2024. Beyond chain-of-thought: A survey of chain-of-x paradigms for llms. *CoRR*.
- Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic chain of thought prompting in large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*. OpenReview.net.
- Ruo Chen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. 2023. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5823–5840. Association for Computational Linguistics.
- Yichun Zhao, Shuheng Zhou, and Huijia Zhu. 2024. Probe then retrieve and reason: Distilling probing and reasoning capabilities into smaller language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13026–13032. ELRA and ICCL.
- Yuhang Zhou and Wei Ai. 2024. Teaching-assistant-in-the-loop: Improving knowledge distillation from imperfect teacher models in low-budget scenarios. In *Findings of the Association for Computational*

Linguistics ACL 2024, pages 265–282. Association for Computational Linguistics.

Xuekai Zhu, Biqing Qi, Kaiyan Zhang, Xinwei Long, Zhouhan Lin, and Bowen Zhou. 2024a. PaD: Program-aided distillation can teach small models reasoning better than chain-of-thought fine-tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2571–2597. Association for Computational Linguistics.

Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. 2024b. Distilling mathematical reasoning capabilities into small language models. *Neural Networks*, 179:106594.

Appendix

A Dataset Statistics

We provide the statistics of the datasets implemented in our study in Table 6, including their original licenses. We follow the train-test dataset splits for GSM8K, ASDiv, SVAMP, and CommonsenseQA from (Chen et al., 2023). For StrategyQA, Tracking Shuffled Objects, Date Understanding, and Last Letter Concatenation, we follow the train-test dataset splits from (Ho et al., 2023).

Meanwhile, in practice, we utilize CoT annotations from (Chen et al., 2023) for GSM8K, ASDiv, SVAMP, CommonsenseQA, and newly prompt the LLM for other datasets. For other datasets, we prompt the LLM of six CoT annotations per question. Furthermore, we report in Table 7 the number of CoT rationales augmented (the size of $\mathcal{D}_{\text{mentor}}$) by our mentor model (FlanT5-large) that has been used in experiments of Section 5.1.

B Implementation Details on Various Student Models

For experiments on models smaller than 1B, we use T5 and FlanT5 as our backbone models with an AdamW optimizer. We conduct a hyperparameter search on τ of {1.0, 1.5, 2.0}, λ of {0.1, 0.2, 0.3, 0.4}, and learning rate of {1e-4, 2e-4, 3e-4, 4e-4, 5e-4}, and report the best test accuracy per epoch. Meanwhile, for labels of the question-label pairs, we adopt the template “{ r_i }. --> { y_i }.” for saving tokenization spaces following (Ho et al., 2023). For experiments on the Vanilla-KD baseline and our Mentor-KD, we randomly select three out of six CoT annotations per question. Moreover, we have the mentor model generate three CoT rationales per question for augmentation.

C Additional Costs for Mentor Models

Method	Train Set	Shuffled	Last Letter
Vanilla-KD	100%	63.11	52.67
Mentor-KD (ours)	100%	82.67	58.67
	80%	82.22	56.00
	40%	67.11	52.67

Table 5: Comparison between Vanilla-KD and Mentor-KD (ours) with different training set ratios.

Although our study mainly spotlights the inference efficiency of small LMs as mentioned in the

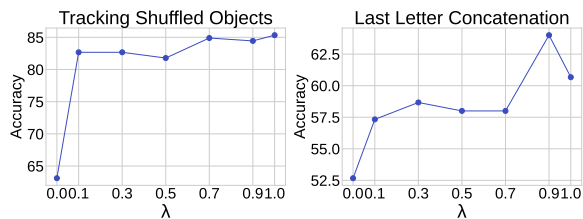


Figure 7: Effects of soft label distillation, by varying the value of loss interpolation hyperparameter (λ).

limitations section, it may be argued that Mentor-KD requires extra computational costs for training the mentor models.

However, considering that Mentor-KD achieves comparable performance with smaller distillation sets from LLM teachers, we suggest that Mentor-KD might be more efficient for training the student models than the baselines. This is especially significant, in regard to the substantial inference cost of LLMs (teacher models) (Ding et al., 2024; Wan et al., 2024). Specifically, Table 5 shows that Mentor-KD works on par, or even exceeds the Vanilla-KD baseline trained on 100% of the distillation sets from the LLM teacher while utilizing only 40% of them (More detailed results are shown in Figure 5). This indicates the potential to save the inference cost of generating 60% of the distillation sets by the LLM teacher. Taking the entire KD pipeline into account, Mentor-KD may train the student more efficiently depending on the design choices, such as the size of the mentor models and the number of distillation sets from the LLM.

D Effects of Soft Label Distillation

In this section, we examine the effects of soft label distillation in Mentor-KD, through differentiating the loss interpolation hyperparameter (λ) in Equation 5. We diversely set the value of λ from 0 (no soft labels) to 1 (only soft labels) in this experiment, and set the student model to FlanT5-small using two reasoning tasks.

The results are shown in Figure 7. We initially observe that the student model’s performances are the lowest when no soft labels are introduced to reasoning distillation. However, we also observe that introducing soft labels significantly contribute to performance boosts of the student, implying that the soft labels which mentor models provide are beneficial to student models carrying out multi-step reasoning.

Dataset	Choices	# Train Data	# Test Data	License	References
StrategyQA	2	1603	687	Apache-2.0	Geva et al. 2021
CommonsenseQA	5	8520	1221	Unspecified	Talmor et al. 2019
ASDiv	-	1462	314	CC BY-NC 4.0	Miao et al. 2020
SVAMP	-	700	150	MIT	Patel et al. 2021
GSM8K	-	7473	659	MIT	Cobbe et al., 2021
Tracking Shuffled Objects	3	525	225	Apache-2.0	Srivastava et al. 2023
Date Understanding	5-6	258	111	Apache-2.0	Srivastava et al. 2023
Last Letter Concatenation	-	350	150	Unspecified	Wei et al. 2022b; Kojima et al. 2022

Table 6: Statistics of datasets used in our study.

Dataset	# Train Data	# Aug Data
StrategyQA	1603	4396
CommonsenseQA	8520	25413
ASDiv	1462	2667
SVAMP	700	1558
Tracking Shuffled Objects	525	1392
Date Understanding	258	763
Last Letter Concatenation	350	1029

Table 7: Statistics of augmented samples by our mentor model (FlanT5-large) per dataset. Here, three CoT rationales per question are generated by our mentor models for augmentation, followed by a filtering process.

E API Usage

As mentioned in Section 4.1, we employ GPT-3.5-Turbo as our teacher model throughout all experiments. Access to the model was provided by the OpenAI API. We set the generation temperature to 0.7, following previous works (Wang et al., 2023b; Ho et al., 2023). Our total expenditure for collecting CoT annotations was \$240.5 USD.

F Case Study

In Table 8, we provide some examples of how our mentor model can successfully augment teacher-incorrect samples on four datasets. Here, we employ GPT-3.5-Turbo as our teacher model, and FlanT5-large as our mentor model. We observe in commonsense reasoning tasks, there are cases where the teacher model does not answer the question faithfully or fails to narrow down its final prediction to a single choice, in contrast to the rationales that the trained mentor model generates. Meanwhile for symbolic reasoning (Last Letter Concatenation), we observe instances where the teacher model makes a final prediction inconsistent of its multi-step rationale, on contrary with the mentor model.

(a) StrategyQA	
Question	Would Jesus understand the Easter Bunny?
Answer	No
Teacher	1. Jesus lived over 2,000 years ago in the Middle East, so it is highly unlikely that he would have encountered the Easter Bunny as it is known in modern Western culture. 2. The Easter Bunny is a symbol of Easter that has its origins in German folklore and was brought to America by German immigrants in the 18th century. It is not a part of Christian religious tradition. 3. Jesus preached about love, forgiveness, and compassion, so it is possible that he would understand and appreciate the idea of the Easter Bunny as a symbol of joy, renewal, and new life. 4. Ultimately, whether or not Jesus would understand the Easter Bunny is a matter of personal interpretation and belief. Some may argue that Jesus would not have a need to understand or acknowledge the Easter Bunny, as his teachings focus on spiritual matters rather than secular traditions. Others may see the Easter Bunny as a harmless and fun part of Easter celebrations. Therefore, the answer is it depends. ✗
Mentor	First, we need to consider the historical context of Easter Bunny. Easter Bunny is a mythical Easter bunny that is associated with the Easter holiday. Jesus was a historical figure who lived during the time of the Roman Empire, during the time of the Roman conquests. Jesus was a man who lived in the first century AD, during the time of the Roman Empire. Therefore, it is highly unlikely that Jesus would have understood the Easter Bunny. -- > No ✓
(b) CommonsenseQA	
Question	What does someone usually produce a lot of when getting divorce? Answer Choices: A. freedom B. tears C. richer D. depression E. tissues
Answer	(B)
Teacher	First, when getting a divorce, someone may produce a lot of tears due to the emotional strain of the process. This would make option B a likely choice. Second, the individual may also experience a sense of freedom from a difficult or unhappy marriage, making option A a possible answer. Third, divorce could potentially lead to financial changes, with the individual becoming either richer or poorer depending on the circumstances. However, it is more common for individuals to experience financial strain during divorce proceedings, making option C less likely. Finally, the emotional toll of divorce can contribute to feelings of depression, making option D a reasonable choice. While tissues may be needed to dry the tears produced during a divorce, they are not typically something that someone produces in large quantities during the process. Therefore, option E is the least likely choice. In conclusion, the most likely answers are B. tears and D. depression. Therefore, the answer is B and D. ✗
Mentor	B. tears When someone is going through a divorce, they are likely to experience emotional turmoil and emotional distress. This can lead to tears as they process the loss and the emotional impact of the situation. --> B ✓
(c) Tracking Shuffled Objects	
Question	Alice, Bob, and Claire are playing a game. At the start of the game, they are each holding a ball: Alice has a red ball, Bob has a pink ball, and Claire has a black ball. the game progresses, pairs of players trade balls. First, Claire and Alice swap balls. Then, Alice and Bob swap balls. Finally, Alice and Claire swap balls. At the end of the game, Alice has the choice is true? Answer choices: (A) red ball, (B) pink ball, (C) black ball.
Answer	(A)
Teacher	After the first swap: Alice has a black ball, Bob has a pink ball, Claire has a red ball. After the second swap: Alice has a red ball, Bob has a pink ball, Claire has a black ball. After the third swap: Alice has a black ball, Bob has a pink ball, Claire has a red ball. Therefore, at the end of the game, Alice has the (C) black ball. Therefore, the answer is (C). ✗
Mentor	After the first swap: Alice has a black ball, Bob has a pink ball, Claire has a red ball. After the second swap: Alice has a pink ball, Bob has a black ball, Claire has a red ball. After the third swap: Alice has a red ball, Bob has a black ball, Claire has a pink ball. Therefore, at the end of the game, Alice has the (A) red ball. --> (A) ✓
(d) Last Letter Concatenation	
Question	Take the last letters of each words in "Dino Toby Abigail Manuela" and concatenate them.
Answer	oyla
Teacher	1. Dino - Last letter: o, 2. Toby - Last letter: y, 3. Abigail - Last letter: l, 4. Manuela - Last letter: a. Concatenate: oyl. Therefore, the answer is oyl. ✗
Mentor	1. Dino - o, 2. Toby - y, 3. Abigail - l, 4. Manuela - a. Concatenating the last letters of each word, we get: oyla. --> oyla ✓

Table 8: Case study of rationale augmentations by the mentor model (FlanT5-large).