Mitigating Spurious Correlations Between Question and Answer via Chain-of-Thought Correctness Perception Distillation

Anonymous ACL submission

Abstract

Large language models (LLMs) excel at reasoning tasks but are expensive to deploy. Thus small language models (SLMs) are fine-tuned on CoT data generated by LLMs to copy LLMs' abilities. However, these CoT data may include noisy rationales that either fail to substantiate the answers or contribute no additional information to support answer prediction, which leads SLMs to capture spurious correlations between questions and answers and compromise the quality of reasoning. In this work, we propose Chain-of-Thought Correctness Perception Distillation (CoPeD), which aims to improve the reasoning quality of the student model from the perspectives of task setting and data utilization. Firstly, we introduce a correctness-aware task setting that encourages the student model to predict answers based on correct rationales and revise them when they are incorrect. This setting improves the faithfulness of reasoning and allows the model to learn from its mistakes. Then, we propose a Correctness-Aware Weighted loss, which dynamically adjusts the contribution of each training instance based on the combined loss of the rationale and the answer. This strategy encourages the model to focus more on samples where the rationale offers stronger support for the correct answer. Experiments have shown that CoPeD is effective on both in-distribution (IND) and out-of-distribution (OOD) benchmark reasoning datasets[1].

1 Introduction

Through progressive scaling of model architectures and training datasets, LLMs have demonstrated exceptional CoT reasoning capabilities in complex NLP tasks. As evidenced by recent studies (Brown et al., 2020; Hoffmann et al., 2022; Chowdhery et al., 2023; OpenAI, 2023; Chen et al., 2023a), the CoT paradigm enables multi-step logical reasoning through explicit intermediate derivations. While this paradigm facilitates complex problem-solving, it also introduces significant computational costs. These costs pose practical challenges for real-world deployment (Shao and Li, 2025). For example, GPT-3 (Brown et al., 2020) has 175 billion parameters. Its inference requires substantial computation, making deployment costly.

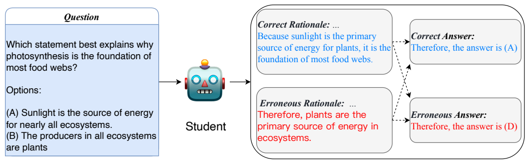

Figure 1: During training, the student model may capture spurious correlations between the question and the answer. As a result, during inference, the rationale could be correct while the answer is erroneous, or the answer could be correct while the rationale is erroneous.

Therefore, the current research (West et al., 2022; Magister et al., 2023; Ho et al., 2023; Fu et al., 2023; Chen et al., 2023b; Zhou and Ai, 2024; Li et al., 2024; Chenglin et al., 2024; Wadhwa et al., 2024; Lee et al., 2024) on knowledge distillation aims to transfer the powerful reasoning ability of LLMs to SLMs. The standard process of this procedure consists of two stages: First, the LLM serves as a teacher to generate rationales for each sample. Then, these rationales are used to perform supervised fine-tuning on the SLM. Although this paradigm improves the reasoning capabilities of SLMs on specific tasks, it commonly assumes that the generated rationale is reliable as long as the predicted answer is correct. However, this assumption does not always hold, as in many cases the rationale neither introduces new information beyond what is provided in the input nor effectively justifies the answer (Wang et al., 2023a). Therefore, during training, SLMs may capture spurious correlations between questions and answers, leading to two main issues, as shown in Figure 1. First, the

[1]We will release our code and data upon publication to facilitate reproducibility.

model may overlook the causal logical relationship between rationales and answers. As a result, the generated rationales may be inconsistent with the predicted answers (Wang et al., 2023a; Feng et al., 2024). Second, such spurious correlations can degrade the quality of rationale generation during reasoning (Dai et al., 2024b). In particular, an error in intermediate steps may lead to error propagation in subsequent reasoning.

To address the above issues, we propose the **C**hain-**o**f-Thought Correctness **Pe**rception **D**istillation (CoPeD). Specifically, we begin by prompting the teacher model (LLM) to generate both correct and erroneous rationales. Following previous work (Dai et al., 2024a,b), we assume that *If the LLM's predicted answer is correct, the rationale is assumed to be correct; otherwise, the rationale is considered erroneous.* Building on this, we introduce a **correctness-aware task setting**. In this setting, the student model predicts the answer based on the question and rationale when the rationale is correct. Otherwise, it generates a revised rationale when the original rationale is erroneous. This task design encourages the student model to rely on valid reasoning paths for answer prediction, thereby mitigating spurious correlations between questions and answers. Since the assumption regarding rationale correctness may not always hold (Wang et al., 2023a), we further propose a **correctness[2]-aware confidence-weighted loss**. This loss dynamically adjusts each sample's contribution to the overall loss by evaluating the degree to which the rationale supports the answer. This mechanism directs the model to focus more on high-quality training examples that demonstrate more reliable reasoning processes and stronger alignment between the rationale and the answer.

Experiments demonstrate that CoPeD outperforms the baselines on both IND and OOD benchmark datasets. Our contributions can be summarized as follows:

- We propose a correctness-aware task setting where the model is trained to answer questions based on correct rationales and to revise erroneous rationales. This design improves the consistency between the generated rationales and answers, thereby enhancing the faithfulness and soundness of reasoning.

- We develop a Correctness-Aware Confidence-Weighted Loss, which jointly considers rationale and answer prediction losses to re-weight training examples. This loss encourages the model to focus more on informative, well-aligned samples, while reducing the impact of noisy or misleading ones.

- We conduct comprehensive experiments across IND and OOD benchmarks, demonstrating that CoPeD effectively improves the reasoning performance of SLMs.

2 Method

The core idea of our method consists of two components: (1) guiding the student model with different training tasks based on the correctness of the rationale; and (2) encouraging the model to prioritize learning from high-quality and logically consistent rationales during training. From the perspectives of task design and data utilization, our approach jointly enhances the faithfulness and soundness of the rationales generated by the student model. The overall framework of our method is illustrated in Figure 2. In this section, we provide a detailed explanation of the method and discuss the motivation behind it.

2.1 Extracting Rationales from Teacher

For each training data sample $\mathcal{D}_{\text{train}} = \{(q_i, a_i)\}_{i=1}^{n}$, we first employ a prompting method to automatically extract correct and erroneous rationales from the teacher model. Specifically, if the LLM's predicted answer matches the ground truth, the corresponding rationale is considered likely correct; otherwise, it is assumed to be erroneous. We collect these rationales for two main purposes: (1) to enable the student model to learn from correct rationales; and (2) to enable the student model to learn how to correct erroneous rationales. This method utilizes a few annotated examples to guide the teacher in generating rationales for new instances (Wei et al., 2022). To maintain the quality of generated CoT, we following Dai et al. (2024a) and use its provided prompt templates to guild the teacher generate correct and erroneous rationales with similar reasoning paths but different conclusions. Eventually, we construct the dataset $\mathcal{D}_{\text{train}} = \{(q_i, r_i^+, r_i^-, a_i)\}_{i=1}^{n}$ for the student model, where q_i is a question, a_i is an answer, r_i^+ is the correct rationale, and r_i^- is the erroneous rationale.

[2]Here, "correctness" refers to whether the rationale provides effective support for the ground-truth answer.

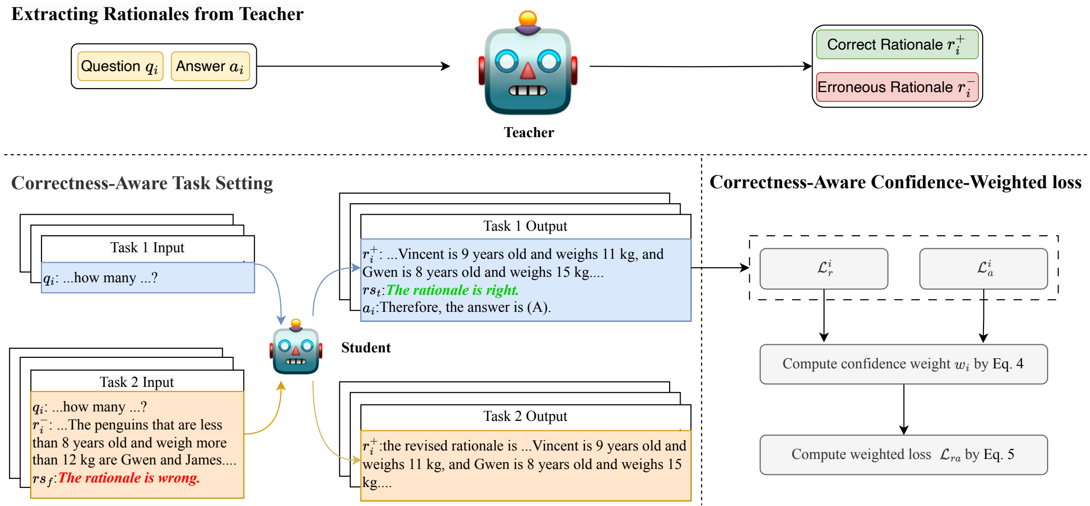

Figure 2: Overview of **Chain-of-Thought Correctness Perception Distillation(CoPeD)**. We use teacher and student models to generate correct and erroneous rationales for the entire training set. Then, we adopt a multi-task learning framework to leverage these rationales, where one task is trained to predict the answer based on correct rationales, and the other task is trained to correct erroneous rationales as additional supervision signals.

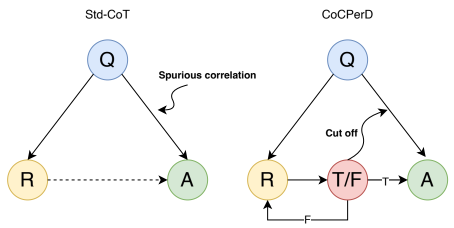

Figure 3: CoPeD adopts different strategies based on the correctness of the rationale, cutting off the spurious correlation between the question and the answer.

2.2 Correctness-Aware Task Setting

To mitigate spurious correlations between questions and answers, we propose a correctness-aware task setting consisting of two tasks: answer prediction and rationale correction. To distinguish between these tasks, we append rationale status tokens rs_t (for correct rationale) and rs_f (for erroneous rationale) to the rationale. When the rationale is correct, the student model predicts the answer based on both the question and the rationale. When the rationale is erroneous, the model learns to revise the rationale. This framework encourages the model to rely on valid reasoning paths for answer prediction, rather than superficial question–answer correlations, thereby enhancing the faithfulness of the generated rationale. Additionally, the rationale correction task helps the model learn from mistakes. This reduces the probability of flawed reasoning steps during inference and improves the soundness of the generated rationale.

In the answer prediction task, the input to the student model is the question q, and its corresponding label consists of three components: the correct rationale r^+, the rationale status string rs_t = **"the rationale is right"**, and the answer a. The loss function for the answer prediction task is formulated as follows:

$$\mathcal{L}_{\mathrm{ra}} = \mathbb{E}_{(q,r^+,a)\sim\mathcal{D}_{\mathtt{train}}}\left[\ell(q, r^+ \oplus rs_\mathrm{t} \oplus a)\right] \quad (1)$$

In the rationale correction task, we concatenate the question q, the erroneous rationale r^- and the rationale status string rs_f = **"the rationale is wrong"** as the input to the student model. The output label is the correct rationale r^+. This task design aims to enable the student model to learn to correct erroneous rationales. It thereby implicitly enhances the student model's robustness and the quality of rationales generated during reasoning. The loss function for the rationale correction task is formulated as follows:

$$\mathcal{L}_{\mathrm{rc}} = \mathbb{E}_{(q,r^+,r^-)\sim\mathcal{D}_{\mathtt{train}}}\left[\ell(q \oplus r^- \oplus rs_\mathrm{f}, r^+)\right] \quad (2)$$

The final objective function jointly optimizes the answer prediction loss $\mathcal{L}$ra and the rationale correction loss $\mathcal{L}$rc, defined as:

$$\mathcal{L}_{\mathrm{CoPeD}} = (1-\alpha)\mathcal{L}_{\mathrm{ra}} + \alpha\mathcal{L}_{\mathrm{rc}}, \quad (3)$$

where α is the hyperparameter used to weight the losses between the two learning tasks.

It is important to note that when rs_f is used as the input or output of the rationale correction task, the student model either determines the correctness of the generated rationale during inference, or it does not, respectively. In practice, we evaluate both inference variants. One variant has the model attempt to assess and revise flawed rationales. The other variant has the model directly generate a rationale without explicit correctness assessment. We find that the performance under both settings is comparable. This suggests that training with the correction task improves the model's ability to generate faithful rationales, even when no correction is applied at inference. We provide a detailed analysis in Appendix D.

2.3 Correctness-Aware Weighted loss

While we initially assign rationale correctness labels based on whether the teacher model's predicted answer matches the ground truth, this heuristic labeling may be noisy in practice. Specifically, a rationale might contain logical flaws despite leading to a correct answer. Conversely, an erroneous answer might be supported by a seemingly plausible rationale. Such coarse-grained supervision can mislead the student model during training. It may cause the model to overfit to unreliable or spurious reasoning paths.

To mitigate this issue, we propose a Correctness-Aware Weighted Loss, which dynamically adjusts the training contribution of each sample based on the degree of alignment between its rationale and answer. This mechanism enables the model to prioritize learning from samples with faithful and consistent rationale–answer pairs. At the same time, it down-weights samples exhibiting reasoning flaws or misalignment. By introducing this label-robust supervision strategy, the student model can better discern and rely on high-quality reasoning paths during distillation.

Concretely, we compute the rationale generation loss $\mathcal{L}_r$ and the answer prediction loss $\mathcal{L}_a$ for each instance. Based on these losses, we calculate a normalized confidence weight w_i, reflecting the reliability of the sample. Specifically, w_i is computed as follows:

$$w_i = \text{softmax}_i\left(-\frac{\mathcal{L}_r^{(i)} + \mathcal{L}_a^{(i)} + |\mathcal{L}_r^{(i)} - \mathcal{L}_a^{(i)}|}{\tau}\right), \quad (4)$$

where composite loss term $\mathcal{L}_r + \mathcal{L}_a$ reflects the overall reliability of a sample. A high rationale loss indicates a noisy reasoning process. In contrast, a high answer prediction loss suggests that the rationale fails to support the final prediction. Furthermore, we use a discrepancy term $|\mathcal{L}_r - \mathcal{L}_a|$ to measure the alignment between the two objectives. The temperature parameter τ controls the smoothness of the resulting weight distribution.

Algorithm 1: Training with Correctness-Aware Weighted loss

Input: Training dataset $\mathcal{D}$, student model f_θ, temperature τ, starting epoch n, cross entropy loss $CE(,)$.

for *epoch* $= 1$ *to* N **do**
  for *each mini-batch* $\mathcal{B} = \{(q_i, a_i, r_i)\}_{i=1}^{B}$ **do**
    for *each sample* $(q_i, a_i, r_i) \in \mathcal{B}$ **do**
      Generate rationale $\hat{r}_i$ and predict answer $\hat{a}_i$: ;
      $\hat{r}_i, \hat{a}_i = f_\theta(q_i)$;
      Compute rationale loss: $\mathcal{L}_r^{(i)} = CE(\hat{r}_i, r_i)$;
      Compute answer loss: $\mathcal{L}_a^{(i)} = CE(\hat{a}_i, a_i)$;
    if *epoch* $< n$ **then**
      Compute unweighted loss: $\mathcal{L}_{ra} = \sum_i(\mathcal{L}_r^{(i)} + \mathcal{L}_a^{(i)})$;
    else
      for *each sample* i *in* $\mathcal{B}$ **do**
        Compute confidence weight w_i by Eq. 4;
      Compute weighted loss $\mathcal{L}_{ra}$ by Eq. 5;
    Update student model parameters using gradient descent;

The Correctness-Aware Weighted loss is then defined as a weighted summation over the sample-wise rationale and answer losses:

$$\mathcal{L}_{ra} = \sum_i w_i \cdot (\mathcal{L}_r^{(i)} + \mathcal{L}_a^{(i)}) \quad (5)$$

To stabilize training, we initially use uniform weights during early epochs to allow the model to acquire basic reasoning capabilities. Starting from epoch n, the correctness-aware weighting mechanism is introduced to emphasize trustworthy samples adaptively. The complete training algorithm is presented in Algorithm 1.

3 Experiments

In this section, we conduct extensive experiments and analyses to evaluate the effectiveness of our method on both in-domain (IND) and out-of-domain (OOD) datasets.

3.1 Datasets

In-domain Dataset: BIG-Bench Hard (BBH) (Suzgun et al., 2023) consists of 27 challenging tasks drawn from BIG-Bench (BB) (Guo et al.), covering domains such as arithmetic, symbolic reasoning, and others. Most tasks are multiple-choice

Method	**Distill?**	**Gen CoT?**	**BBH-test**	**BB-sub**	**AGIEval**	**ARC-E**	**ARC-C**	**AVG**
In-domain?			✓	×	×	×	×	
			Teacher: ChatGPT (gpt-3.5-turbo)					
Zero-shot-CoT	×	✓	42.6	44.5	50.3	92.1	82.2	62.3
			Student: LLaMA2-7B					
Std-CoT (Magister et al., 2023)	✓	✓	58.5	29.5	24.2	61.8	47.3	44.3
SCOTT (Wang et al., 2023a)	✓	✓	43.1	19.7	12.8	46.3	35.9	31.6
MT-CoT (Li et al., 2022)	✓	✓	59.3	31.4	23.2	51.7	40.6	41.2
Step-by-step (Hsieh et al., 2023)	✓	✓	44.6	29.2	28.4	69.0	49.2	43.2
CasCoD (Dai et al., 2024b)	✓	✓	60.2	37.2	28.6	<u>71.1</u>	<u>52.4</u>	49.9
CoPeD-T (ours)	✓	✓	<u>63.1</u>	<u>38.3</u>	<u>30.2</u>	**72.6**	**55.1**	<u>51.8</u>
CoPeD-L (ours)	✓	✓	60.9	38.2	27.9	69.2	50.9	49.4
CoPeD-TL (ours)	✓	✓	**69.8**	**39.5**	**31.7**	<u>71.1</u>	52.3	**52.9**
			Student: Mistral-7B-v0.2					
Std-CoT(Magister et al., 2023)	✓	✓	72.5	36.8	32.5	67.6	58.6	53.6
SCOTT (Wang et al., 2023a)	✓	✓	31.9	32.8	27.3	54.2	38.6	37.0
MT-CoT (Li et al., 2022)	✓	✓	56.1	39.4	31.3	68.4	59.3	50.9
Step-by-step(Hsieh et al., 2023)	✓	✓	58.1	37.5	22.9	78.4	61.7	51.7
CasCoD(Dai et al., 2024b)	✓	✓	70.5	39.5	<u>38.2</u>	**84.2**	**75.5**	**61.6**
CoPeD-T (ours)	✓	✓	<u>74.4</u>	40.7	36.7	78.9	68.5	59.8
CoPeD-L (ours)	✓	✓	74.1	<u>40.8</u>	36.1	80.1	65.4	59.3
CoPeD-TL (ours)	✓	✓	**75.2**	**41.2**	**38.5**	<u>82.6</u>	<u>68.6</u>	<u>61.2</u>

Table 1: Accuracy (%) on in-domain and out-of-domain datasets with different methods. ♠: the results borrowed from Dai et al. (2024b). The best performance among distilled student models is marked in **bold**, and the second-best performance is indicated by an <u>underline</u>. CoPeD-T denotes the correctness-aware task setting, CoPeD-L refers to training Std-CoT using a correctness-aware weighted loss, and CoPeD-TL represents the combination of both methods.

questions, with a few open-ended ones. Following Dai et al. (2024b), we randomly split the BBH dataset into a training set (BBH-train) for distillation and a test set (BBH-test) for IND evaluation, using a 4:1 split.

Out-of-domain Dataset: **(1) BIG-Bench Sub (BB-sub)** is derived from BIG-Bench (BB) (Guo et al.), encompassing 203 tasks across domains such as linguistics, mathematics, and commonsense reasoning. To simplify our evaluation, we use the BB-Sub filtered by Dai et al. (2024b). **(2) AGIEval** (Zhong et al., 2024) is a benchmark that assesses language models (LMs) on reasoning abilities using human exams from fields including English, Mathematics, Law, and Logic. We select the English multiple-choice question subtask filtered by Dai et al. (2024b). **(3) AI2 Reasoning Challenge (ARC)** (Clark et al., 2018) consists of two datasets: ARC-Easy and ARC-Challenge, derived from middle and high school science exams. ARC-E features easier questions, while ARC-C presents more challenging ones. Following Dai et al. (2024b), we use the test sets from both datasets for evaluation.

3.2 Implementation Details

Models We use LLaMA-7B (Touvron et al., 2023) as the base student model throughout all experiments unless otherwise specified. Given its cost-effectiveness and capabilities, we leverage OpenAI's powerful black-box LLM, gpt-3.5-turbo-0613, as the teacher model to extract chain-of-thoughts (CoTs) using the same manual prompt as in prior works (Dai et al., 2024a).

Setup We use LoRA (Hu et al.) for parameter-efficient fine-tuning of the student model. To balance the answer prediction and rationale correction tasks, we set α to 0.5. All experiments are performed using a mixed-precision training strategy on 8 × A800 GPUs. During inference, we utilize vLLM3 (Kwon et al., 2023) to accelerate the process, employing a greedy decoding strategy for text generation on a single A800 GPU. Further details on training and hyperparameters are provided in Appendix B.1.

Baselines We compare our method with the following baselines: (1) **Teacher** in Zero-shot-CoT(Kojima et al., 2022) for showing the impact of distilling reasoning ability from LLMs. (2) **Std-CoT** (Magister et al., 2023), which is the standard CoTs distillation method that directly fine-tune student models on the CoTs data. (3) **MT-CoT** (Li et al., 2022) is also a multi-task CoTs distillation method, but unlike Step-by-step, it simultaneously optimizes the objectives of answer prediction and entire CoTs learning. (4) **Step-by-step** (Hsieh et al., 2023) is a multi-task CoTs distillation method that distills rationales and answers separately (5) **SCOTT** (Wang et al., 2023a) that enhances the reasoning consistency of the student model by introducing additional counterfactual data. (6) **CasCoD** (Dai et al., 2024b) splitting single-step learning into two cascaded steps, restructuring training objectives to enhancing reasoning generalizability.

3.3 Main Results

As shown in Table 1, CoPeD demonstrates competitive performance against strong baselines on both in-domain (IND) and out-of-domain (OOD) benchmarks. Specifically, LLaMA2-7B equipped with CoPeD-TL achieves an average accuracy of 52.9% across all tasks. It outperforms the strongest baseline, CasCoD, by 3.0%. In particular, it surpasses CasCoD by 9.6% in the IND scenario. Meanwhile, CoPeD-TL also exhibits strong generalization ability in OOD scenarios. It outperforms CasCoD by 2.3%, 3.1%, 1.5%, and 2.7% on BB-test, AGIEval, ARC-E, and ARC-C, respectively. While Mistral-7B with CoPeD-TL does not achieve the best results on every individual OOD scenario, it still delivers consistently competitive performance across a wide range of tasks. This highlights the robustness and generalization ability of our method, even in comparison with state-of-the-art models.

These results validate the effectiveness of our design. CoPeD-T enables the student model to discern faithful reasoning paths from spurious ones, thereby improving both the quality and reliability of reasoning. Meanwhile, compared to the Std-CoT approach, CoPeD-L enhances training robustness by adaptively down-weighting misaligned or unreliable samples. It emphasizes high-quality, well-aligned rationales, allowing the model to focus on trustworthy reasoning paths.

3.4 Faithfulness and Soundness of Students

Inspired by previous work (Wang et al., 2023a; Dai et al., 2024b), we employ LLMs as evaluators to assess two aspects. First, whether the rationale provided by the student model supports its prediction (i.e., faithfulness). Second, whether the rationale supports the ground-truth answer (i.e., soundness). Given a rationale $\hat{r}_i$ generated by the student model and an answer (either the predicted answer $\hat{a}_i$ or the ground-truth answer a_i), we construct evaluation prompt[3] p_e to guide LLM-based scoring. We define faithfulness and soundness as follows:

$$\text{Faithfulness} = \mathbb{E}\big[f_{\text{eval}}(p_e, q_i, \hat{r}_i, \hat{a}_i)\big], \quad (6)$$

$$\text{Soundness} = \mathbb{E}\big[f_{\text{eval}}(p_e, q_i, \hat{r}_i, a_i)\big], \quad (7)$$

where $f_{\text{eval}}(\hat{r}_i, \hat{a}_i)$ and $f_{\text{eval}}(\hat{r}_i, a_i) \in \{0, 1\}$ are a binary evaluation function, returning 1 if the rationale $\hat{r}_i$ sufficiently supports the given answer (either the predicted answer $\hat{a}_i$ or the ground-truth answer a_i), and 0 otherwise.

Method	Faithfulness			Soundness		
	ChatGPT	GPT4	AVG	ChatGPT	GPT4	AVG
Teacher	86.6	86.9	86.8	74.8	71.5	73.2
Std-CoT	80.5	67.9	74.2	64.0	54.5	59.3
CasCoD	82.2	72.6	77.4	70.2	59.6	64.9
CoPeD-TL	83.8	78.5	81.2	72.6	67.9	70.2

Table 2: Faithfulness (%) and Soundness (%) of the compared methods on the IND dataset. We employ both ChatGPT and GPT-4 as evaluators to mitigate the risk of single-model bias.

The results are shown in Table 2. Compared to the baseline, the rationale generated by CoPeD-TL is more consistent with the answer. This includes both the predicted and the ground-truth answers. This indicates that CoPeD-TL ensures the faithfulness and soundness of the rationale generated during the reasoning process. It does so by adopting different strategies based on the correctness of the rationale and filtering noisy samples. This approach helps mitigate spurious correlations between the question and the answer.

3.5 Ablation Study

Model Size We conducted model distillation on TinyLLaMA-1.1B[4] (Zhang et al., 2024), LLaMA2-7B, and LLaMA2-13B, and compared it with Std-CoT, MT-CoT, and CasCod. As shown in Figure 4,

[3]The prompt for evaluating whether the rationale provided by the student model supports the answer can be found in the Appendix C.2

[4]https://huggingface.co/TinyLlama/TinyLlama-1.1B-intermediate-step-1431k-3T

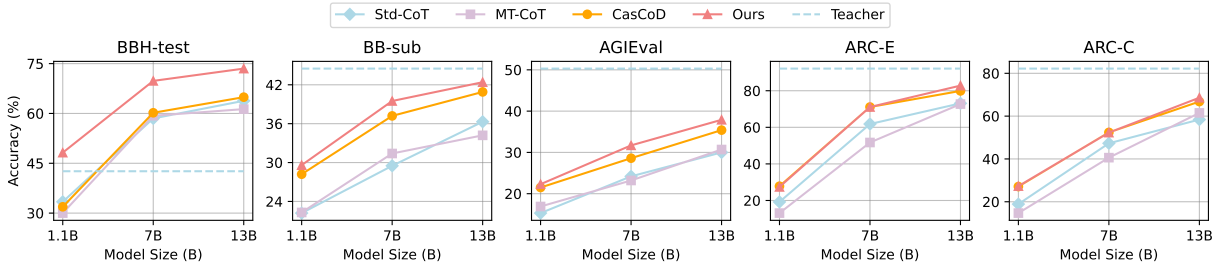

Figure 4: Ablation study on model size for IND and four OOD datasets. The dotted line indicates the performance of the teacher LLM under the Zero-shot-CoT setting.

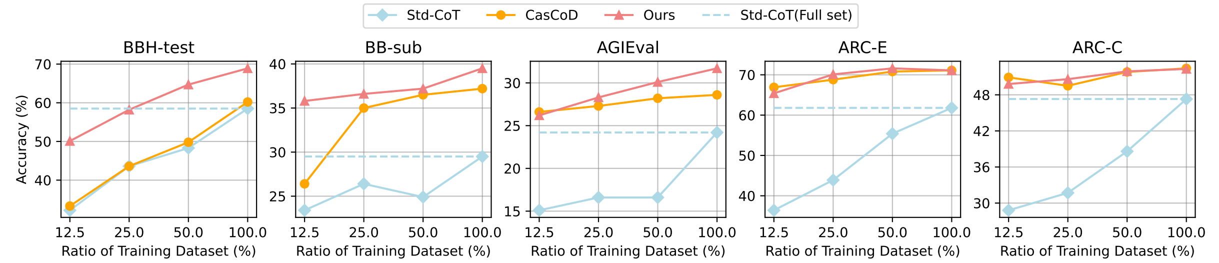

Figure 5: Ablation study on training data size for IND and four OOD datasets. The dotted line indicates the performance of fine-tuning the student models by Std-CoTs distillation using the full set (100% of) BBH-train dataset.

CoPeD-TL consistently achieves competitive performance across student models of varying sizes, outperforming baseline methods on both IND and OOD datasets. Notably, on the IND dataset, the 1.1B model with CoPeD-TL reaches 113.1% of the teacher model's performance, demonstrating the significant advantages of CoPeD-TL in low-resource scenarios. Moreover, across different model scales, CoPeD-TL maintains competitive performance on OOD datasets compared to baseline approaches.

Data Size CoPeD-TL demonstrates significant improvements over baseline methods on both IND and OOD datasets, while utilizing considerably less training data. As shown in Figure 5, CoPeD-TL achieves a 16.8% improvement over CasCoD on the IND (BBH-test) dataset, using only 12.5% of the full BBH-train data. The performance on OOD datasets is even more notable. For example, on the BB-sub dataset, CoPeD-TL achieves a 9.4% improvement in accuracy compared to CasCoD, even when using only 12.5% of the complete BBH-train data. On other OOD datasets, CoPeD-TL also achieves excellent performance. These results clearly demonstrate the effectiveness of CoPeD-TL in low-resource settings. They highlight its ability to enhance the performance of CoTs in both IND and OOD scenarios while requiring significantly less training data.

Method	**Accuracy (%)**	**Gain (%)**
Uniform Weight (Std-CoT)	58.5	–
Only Composite Term ($\mathcal{L}_r + \mathcal{L}_a$)	60.2	+1.7
Only Discrepancy Term ($\lvert\mathcal{L}_r - \mathcal{L}_a\rvert$)	59.8	+1.3
Full (CoPeD-L)	**60.9**	**+2.4**

Table 3: Ablation results of different loss weighting strategies on BBH-test.

Loss Term To evaluate the individual contributions of the components within our correctness-aware weighted loss, we conduct an ablation study focusing on the BBH-test accuracy, as shown in Table 3. When only the composite loss term $(\mathcal{L}_r + \mathcal{L}_a)$ is used to compute sample weights, the model achieves a moderate improvement over the uniform weighting baseline. This suggests that considering the overall sample difficulty helps filter out noisy examples. Similarly, utilizing only the discrepancy term $|\mathcal{L}_r - \mathcal{L}_a|$ yields a modest accuracy gain. This indicates that rationale–answer alignment alone also provides useful supervision. Importantly, combining both components in the full weighting scheme leads to a substantial boost in performance. These results highlight the importance of modeling both sample quality and rationale–answer consistency, enabling the student model to focus on more trustworthy reasoning trajectories.

3.6 Analysis & Case Study

Due to page limitations, we provide analysis in Appendix E and case study in Appendix F.

4 Related Works

Chain-of-Thought Distillation Recent studies have demonstrated that CoT prompts significantly enhance the reasoning ability of LLMs for complex tasks (Wei et al., 2022; Kojima et al., 2022; Wang et al.; Huang et al., 2023). However, this advantage is most pronounced in LLMs, prompting several researchers (Magister et al., 2023; Ho et al., 2023; Li et al., 2023; Chae et al., 2023; Yang et al., 2024) to explore methods for transferring reasoning knowledge from LLMs to SLMs. Typically, these approaches leverage CoT prompts to generate rationales from LLMs, which are then used to fine-tune SLMs.

In addition, Hsieh et al. (2023) argue that reasoning bases and answers should be treated as distinct optimization objectives. Similarly, Li et al. (2022) suggest that learning both the complete CoT and individual answers can enhance the reasoning capabilities of the student model. Liu et al. (2024) introduce an additional distillation objective focused on self-assessment, enabling the SLM to evaluate the accuracy of its generated CoTs. Wang et al. (2023a) propose reducing reasoning errors and hallucinations inherited by the SLM from the LLM through contrastive decoding, which ensures that the reasoning basis is closely related to the answer. Moreover, Wang et al. (2023b) present an interactive, multi-turn paradigm that allows the SLM to engage in self-reflection and receive feedback from the LLM during the learning process. Dai et al. (2024b) suggest decomposing the traditional single-step learning process into two cascading steps to alleviate the effects of spurious correlations between questions and answers. Lee et al. (2024) effectively enhances the reasoning ability of small models by introducing an intermediate-sized, task-specific "mentor" model to improve the quality of multi-step reasoning distillation and provide soft labels. Feng et al. (2024) proposes a counterfactual distillation framework that improves the reasoning ability and OOD robustness of small language models. Wadhwa et al. (2024) investigates why CoT rationales help in model distillation and finds that even incoherent or partial rationales appended after labels can significantly improve student model performance. Chenglin et al. (2024) proposed a Mixed Distillation framework that efficiently distills multi-step reasoning abilities into small models through multi-path reasoning samples and multi-task loss. Zhu et al. (2024) proposes Program-aided Distillation (PaD), which improves the distillation quality of reasoning tasks by using reasoning programs to correct errors in synthetic data and iteratively refine the distilled model's reasoning capabilities.

Learning from Mistakes Recent studies have investigated the use of mistake data to improve the performance of language models. Shinn et al. (2024) introduce Reflexion, a method that allows LLM agents to self-reflect on their mistakes. Wang and Li (2023) propose a study assistant that collects and retrieves training mistakes from LLMs to guide future inferences. Li et al. present the CoK method, which corrects reasoning errors by retrieving relevant knowledge to prevent the propagation of errors. However, these approaches are not directly applicable to standard SLMs. Wang et al. (2023a) propose fine-tuning on counterfactual data to ensure the faithful reasoning of the student model. An et al. (2023) introduce LEMA, a method that fine-tunes language models on corrected mistake data, with mistakes collected from various LLMs. Tong et al. (2024) explores whether large language models (LLMs) can enhance their reasoning abilities by learning from their mistakes, proposing two methods—self-rethinking prompting and mistake tuning. An et al. (2024) investigates whether large language models (LLMs) can enhance their CoT reasoning by learning from mistakes.

5 Conclusion

In this study, we propose a Chain-of-Thought Correctness Perception Distillation framework (CoPeD). It employs a dual-task training mechanism comprising answer prediction and rationale correction to significantly enhance the faithfulness and soundness of reasoning. To address the noise present in teacher-generated data, we introduce a correctness-aware weighted loss. This loss effectively reduces the negative impact of unreliable samples and strengthens the model's ability to identify and leverage high-quality reasoning paths. Extensive experiments across varying model sizes and training data volumes demonstrate that CoPeD consistently achieves superior performance on both IND and OOD benchmarks, validating its effectiveness in improving reasoning quality and generalization capability.

6 Limitations

In our study, we explore enabling the student model to verify the correctness of the generated rationale during inference and to attempt corrections when the rationale is identified as erroneous. However, the student model currently struggles to effectively validate whether the rationale derived from its reasoning is indeed correct.

Moreover, our approach depends on correctness labels heuristically derived from whether the predicted answer matches the ground truth. Although our weighted loss function alleviates the impact of this coarse supervision, it may still introduce noise in certain edge cases. Additionally, the current weighting mechanism is relatively simple and could be further refined to better capture subtle inconsistencies between rationales and answers.

References

Shengnan An, Zexiong Ma, Siqi Cai, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. 2024. Can llms learn from mistakes? an empirical study on reasoning tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 833–854.

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. 2023. Learning from mistakes makes llm better reasoner. *arXiv preprint arXiv:2310.20689*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Hyungjoo Chae, Yongho Song, Kai Ong, Taeyoon Kwon, Minjin Kim, Youngjae Yu, Dongha Lee, Dongyeop Kang, and Jinyoung Yeo. 2023. Dialogue chain-of-thought distillation for commonsense-aware conversational agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5606–5632.

Hongzhan Chen, Siyue Wu, Xiaojun Quan, Rui Wang, Ming Yan, and Ji Zhang. 2023a. Mcc-kd: Multi-cot consistent knowledge distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6805–6820.

Zeming Chen, Qiyue Gao, Antoine Bosselut, Ashish Sabharwal, and Kyle Richardson. 2023b. Disco: Distilling counterfactuals with large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5514–5528.

Li Chenglin, Qianglong Chen, Liangyue Li, Caiyu Wang, Feng Tao, Yicheng Li, Zulong Chen, and Yin Zhang. 2024. Mixed distillation helps smaller language models reason better. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1673–1690.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457.

Chengwei Dai, Kun Li, Wei Zhou, and Songlin Hu. 2024a. Beyond imitation: Learning key reasoning steps from dual chain-of-thoughts in reasoning distillation. *arXiv preprint arXiv:2405.19737*.

Chengwei Dai, Kun Li, Wei Zhou, and Songlin Hu. 2024b. Improve student's reasoning generalizability through cascading decomposed cots distillation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15623–15643.

Tao Feng, Yicheng Li, Li Chenglin, Hao Chen, Fei Yu, and Yin Zhang. 2024. Teaching small language models reasoning through counterfactual distillation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5831–5842.

Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023. Specializing smaller language models towards multi-step reasoning. In *Proceedings of the 40th International Conference on Machine Learning*, pages 10421–10430.

Geyang Guo, Ranchi Zhao, Tianyi Tang, Xin Zhao, and Ji-Rong Wen. Beyond imitation: Leveraging fine-grained quality signals for alignment. In *The Twelfth International Conference on Learning Representations*.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14852–14882.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 30016–30030.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2023. Large language models can self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1068.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, et al. 2024. Training language models to self-correct via reinforcement learning. *arXiv preprint arXiv:2409.12917*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Hojae Lee, Junho Kim, and SangKeun Lee. 2024. Mentor-kd: Making small language models better multi-step reasoners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17643–17658.

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 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2665–2679.

Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, et al. 2022. Explanations from large language models make small reasoners better. *arXiv preprint arXiv:2210.06726*.

Xiang Li, Shizhu He, Jiayu Wu, Zhao Yang, Yao Xu, Yang jun Jun, Haifeng Liu, Kang Liu, and Jun Zhao. 2024. Mode-cotd: Chain-of-thought distillation for complex reasoning tasks with mixture of decoupled lora-experts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11475–11485.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Shafiq Joty, Soujanya Poria, and Lidong Bing. Chain-of-knowledge: Grounding large language models via dynamic knowledge adapting over heterogeneous sources. In *The Twelfth International Conference on Learning Representations*.

Weize Liu, Guocong Li, Kai Zhang, Bang Du, Qiyuan Chen, Xuming Hu, Hongxia Xu, Jintai Chen, and Jian Wu. 2024. Mind's mirror: Distilling self-evaluation capability and comprehensive thinking from large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6748–6763.

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.

R OpenAI. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5).

Jiawei Shao and Xuelong Li. 2025. Ai flow at the network edge. *IEEE Network*, pages 1–1.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, et al. 2023. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051.

Yongqi Tong, Dawei Li, Sizhe Wang, Yujia Wang, Fei Teng, and Jingbo Shang. 2024. Can llms learn from previous mistakes? investigating llms' errors to boost for reasoning. *arXiv preprint arXiv:2403.20046*.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai

Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melissa Hall Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Somin Wadhwa, Silvio Amir, and Byron C Wallace. 2024. Investigating mysteries of cot-augmented distillation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6071–6086.

Danqing Wang and Lei Li. 2023. Learning from mistakes via cooperative study assistant for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10667–10685.

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.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Zhaoyang Wang, Shaohan Huang, Yuxuan Liu, Jiahai Wang, Minghui Song, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, et al. 2023b. Democratizing reasoning ability: Tailored learning from large language model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1948–1966.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625.

Bohao Yang, Chen Tang, Kun Zhao, Chenghao Xiao, and Chenghua Lin. 2024. Effective distillation of table-based reasoning ability from llms. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5538–5550.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. Agieval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314.

Yuhang Zhou and Wei Ai. 2024. Teaching-assistant-in-the-loop: Improving knowledge distillation from imperfect teacher models in low-budget scenarios. *arXiv preprint arXiv:2406.05322*.

Xuekai Zhu, Biqing Qi, Kaiyan Zhang, Xinwei Long, Zhouhan Lin, and Bowen Zhou. 2024. 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.

A Additional Experiment Results

A.1 Ablation on Temperature τ

Temperature τ	1	2.5	5.0	7.5	10.0
BBH-test Accuracy (%)	64.8	69.2	**69.8**	68.5	66.1

Table 4: Effect of temperature τ in the correctness-aware weighting loss on BBH-test accuracy.

We conduct an ablation study to investigate the impact of the temperature coefficient τ in our proposed CoPeD-TL. As shown in Table 4, we evaluate CoPeD-TLon the BBH-test set under various τ values. The results show that performance increases as τ grows from 0.5 to 2.5, peaking at 69.8%. Further increasing τ causes a slight decline, indicating a trade-off between weighting sharpness and generalization. Specifically, small τ values lead to overconfident weighting on a few low-loss samples, which may introduce bias and hinder learning from diverse correct examples. Conversely, large τ makes the softmax distribution nearly flat, diminishing the benefit of correctness-aware weighting.

A.2 Ablation on Task Balance Parameter α

To investigate the impact of task balancing between rationale correction and answer prediction, we conduct an ablation study on the weighting parameter α used in our multi-objective loss. Specifically, we vary $\alpha \in 0.3, 0.5, 0.7$ to control the relative importance of rationale supervision versus answer supervision. The results are reported in Table 5.

From the results, we observe that $\alpha = 0.5$ yields the best average performance across all benchmark datasets, indicating that a balanced emphasis on both rationale correction and answer prediction leads to more effective reasoning. When α is set to 0.3, the model prioritizes answer prediction while underutilizing the benefits of rationale supervision, resulting in suboptimal generalization. Conversely, when α is increased to 0.7, the model places excessive focus on rationale correction, potentially at the cost of answer prediction quality. These findings suggest that maintaining a moderate balance between the two tasks is crucial for achieving strong overall reasoning ability and robustness across diverse benchmarks. Therefore, we adopt $\alpha = 0.5$ as the default setting in our final model.

B Experimantal Settings

B.1 Hyperparameters Settings

To guarantee the fairness of our comparative analysis, in our study, we keep the hyperparameter settings consistent across all baselines, our proposed CoPeD approach included. Below, we provide a detailed account of the hyperparameter configurations used in our experiments. The detailed hyperparameters in training and inference can be found in Table 6 and Table 7, respectively.

In our research, We maintain a consistent batch size across all baselines to eliminate performance differences caused by varying batch sizes. Through a series of experiments with learning rates set to 5e-5, 1e-4, 2e-4, 3e-4 and 4e-4, 5e-4 we find that the learning rate is a critical factor affecting model performance and that the optimal value varies with model size. Therefore, we adjust the learning rate accordingly based on model size.

B.2 Dataset Statistics

Table 8, Table 9, Table 10 and Table 11 show the data statistics of AGIEval, ARC, BIG-Bench Hard (BBH) and BIG-Bench Sub (BB-sub), respectively.

C Prompts

C.1 Prompts of Correct the Erroneous Rationale for ChatGPT

We use the prompt template shown in Table 12 to call the ChatGPT API to correct the erroneous rationale of student model for the BBH-train datasets.

C.2 Prompts of Evaluator

We use the prompt templates shown in Table 13 to call the ChatGPT and GPT-4 APIs, predicting whether the rationale supports the answer.

D Inference Process

Figure 6 demonstrates that different training methods lead to variations in the student model's ability to verify the correctness of the rationale during inference. When the rationale status string rs_{f} is used as the model's input in the rationale correction task, the student model cannot predict the correctness of the rationale during inference. However, when rs_{f} is used as the model's target output in the task, the student model can predict the correctness of the generated rationale during inference and adopt

α	BBH-test	BB-sub	AGIEval	ARC-E	ARC-C	AVG
0.3	70.4	37.7	28.1	69.7	51.4	51.5
0.5	69.8	39.5	31.7	71.1	52.3	**52.9**
0.7	63.5	38.6	29.7	70.2	50.7	50.5

Table 5: Ablation study on the task balance parameter α. α controls the trade-off between answer prediction and rationale correction.

Hyperparameter	**TinyLLaMA-1.1B**	**LLaMA2-7B**	**LLaMA2-13B**	**Mistral-7B-v0.2**
gradient accumulation steps	2	2	2	2
per device batch size	2	2	2	2
learning rate	4e-4	3e-4	1e-4	3e-4
epoches	30	20	20	20
temperature τ	2.5	5	5	5
starting epoch n	10	5	5	5
max length	1024	1024	1024	1024
β of AdamW	(0.9,0.999)	(0.9,0.999)	(0.9,0.999)	(0.9,0.999)
ϵ of AdamW	1e-8	1e-8	1e-8	1e-8
γ of Scheduler	0.95	0.95	0.95	0.95
weight decay	0	0	0	0
warmup ratio	0	0	0	0
rank of LoRA	64	64	64	64
α of LoRA	32	32	32	32
target modules	q_proj, v_proj	q_proj, v_proj	q_proj, v_proj	q_proj, v_proj
drop out of LoRA	0.05	0.05	0.05	0.05

Table 6: Training hyperparameters.

Arguments	**Student**	**Teacher**
do sample	False	True
temperature	-	0.2
top-p	1.0	1.0
top-k	-	-
max new tokens	1024	2048
# return sequences	1	1

Table 7: Generation configs of students and teachers.

Task	**Size**	**# Choices**
ARC-E	2376	4-5
ARC-C	1172	4-5

Table 9: Statistics of ARC test dataset.

No.	**Task**	**Size**	**# Choices**
1	AQuA-RAT	254	5
2	LogiQA-EN	651	4
3	LSAT-AR	230	5
4	LSAT-LR	510	5
5	LSAT-RC	269	5
6	SAT-Math	220	4
7	SAT-EN	206	4
8	SAT-EN (w/o Psg.)	206	4
	Sum	2546	-

Table 8: Statistics of AGIEval dataset.

different strategies based on its correctness. If the student model predicts the generated rationale is correct, it directly predicts the answer based on that rationale. If the rationale is predicted to be erroneous, the model first corrects the rationale, and then uses the corrected rationale along with the question and the verification string rs_{t} as input to predict the answer. When the rationale status string rs_{f} is used as the target output in the rationale correction task, although the student model can verify whether the generated rationale is correct during inference, it still struggles with effectively validating the correctness of the rationale (Kumar et al., 2024). Therefore, our goal is to improve the quality of the rationales generated by the student model during inference by enabling the student model to learn from errors through a rationale correction task, rather than validating the correctness of the generated rationale during inference. In §E, we further discuss the specific impact on student model performance when the rationale status string rs_{f} is used as the student model's input and output in the rationale correction task, respectively.

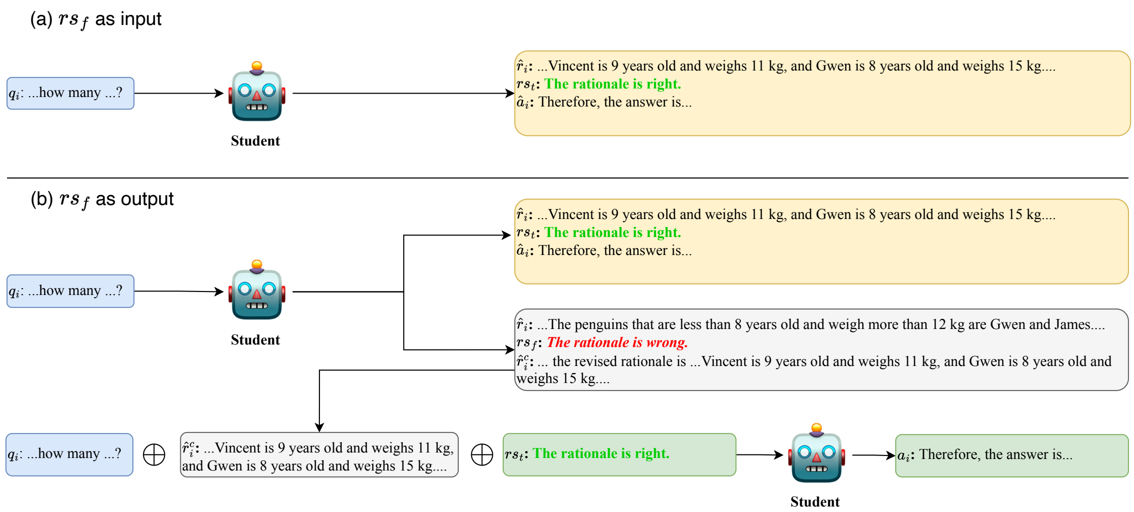

Figure 6: Comparison of student model inference processes under different training strategies.

No.	Task	Size	# Choices
1	Reasoning about Colored Objects	250	18
2	Geometric Shapes	250	11
3	Ruin Names	250	11
4	Penguins in a Table	146	5
5	Movie Recommendation	250	5
6	Tracking Shuffled Objects (3 objects)	250	3
7	Tracking Shuffled Objects (5 objects)	250	5
8	Tracking Shuffled Objects (7 objects)	250	7
9	Logical Deduction (3 objects)	250	3
10	Logical Deduction (5 objects)	250	5
11	Logical Deduction (7 objects)	250	7
12	Date Understanding	250	6
13	Salient Translation Error Detection	250	6
14	Causal Judgement	187	2
15	Disambiguation QA	250	4
16	Temporal Sequences	250	4
17	Boolean Expressions	250	2
18	Hyperbaton (Adjective Ordering)	250	2
19	Navigate	250	2
20	Snarks	178	2
21	Sports Understanding	250	2
22	Formal Fallacies Syllogisms Negation	250	2
23	Web of Lies	250	2
24	Dyck Languages	250	-
25	Multi-Step Arithmetic	250	-
26	Object Counting	250	-
27	Word Sorting	250	-
	Sum	6511	-

Table 10: Statistics of BIG-Bench Hard dataset.

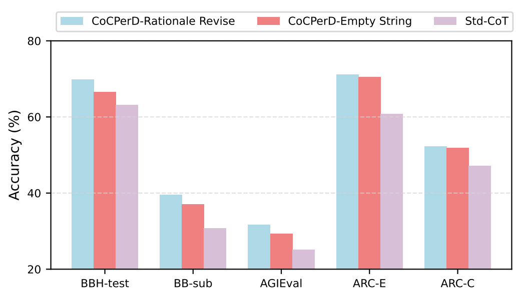

Figure 7: Compare training CoPeD with different target outputs when the rationale is erroneous.

E Analysis

What is the impact of training the student model with different target outputs when the rationale is erroneous? We investigate the impact of training the student model to adopt different target outputs when the rationale is erroneous. As shown in Figure 7, the performance of the student model trained with an empty string as the target output when a reasoning error occurs is significantly lower than that of the student model trained with the correct rationale as the target. This suggests that the rationale correction task implicitly improves the quality of the rationales generated by the student model. Furthermore, the performance of the student model trained with an empty string as the target output is notably superior to that of Std-CoT, which further demonstrates that CoPeD-TL enables the student model to benefit from the generated rationale when predicting answers, thereby effectively mitigating the spurious correlation between the question and the answer.

Whether the student model can effectively verify the correctness of the rationale? We explore the impact of using the rationale status string rs_f as both input and output on CoPeD-TL's performance in the rationale correction task on IND and OOD datasets. As shown in Figure 8, the experiment includes the following three settings: (1) input: When the rationale status string rs_f is used as input, the student model predicts the answer based on the generated rationale without verifying the correctness of the rationale; (2) output-correction: When the rationale status string rs_f is used as output, the student model, after identifying rationale errors, corrects

No.	Task	Size	# Choices
1	abstract_narrative_understanding	100	5
2	anachronisms	100	2
3	analogical_similarity	100	7
4	analytic_entailment	70	2
5	cause_and_effect	100	2
6	checkmate_in_one	100	26
7	cifar10_classification	100	10
8	code_line_description	60	4
9	conceptual_combinations	100	4
10	crass_ai	44	4
11	elementary_math_qa	100	5
12	emoji_movie	100	5
13	empirical_judgments	99	3
14	english_russian_proverbs	80	4
15	entailed_polarity	100	2
16	entailed_polarity_hindi	100	2
17	epistemic_reasoning	100	2
18	evaluating_information_essentiality	68	5
19	fantasy_reasoning	100	2
20	figure_of_speech_detection	59	10
21	goal_step_wikihow	100	4
22	gre_reading_comprehension	31	5
23	human_organs_senses	42	4
24	identify_math_theorems	53	4
25	identify_odd_metaphor	47	5
26	implicatures	100	2
27	implicit_relations	82	25
28	indic_cause_and_effect	100	2
29	intersect_geometry	100	26
30	kanji_ascii	100	5
31	kannada	100	4
32	key_value_maps	100	2
33	logic_grid_puzzle	100	3
34	logical_args	32	5
35	logical_fallacy_detection	100	2
36	metaphor_boolean	100	2
37	metaphor_understanding	100	4
38	minute_mysteries_qa	100	4
39	mnist_ascii	100	10
40	moral_permissibility	100	2
41	movie_dialog_same_or_different	100	2
42	nonsense_words_grammar	50	4
43	odd_one_out	86	5
44	parsinlu_qa	100	4
45	physical_intuition	81	4
46	play_dialog_same_or_different	100	2
47	presuppositions_as_nli	100	3
48	riddle_sense	49	5
49	similarities_abstraction	76	4
50	simple_ethical_questions	100	4
51	social_iqa	100	3
52	strange_stories	100	2
53	strategyqa	100	2
54	swahili_english_proverbs	100	4
55	swedish_to_german_proverbs	72	4
56	symbol_interpretation	100	5
57	timedial	100	3
58	undo_permutation	100	5
59	unit_interpretation	100	5
60	vitaminc_fact_verification	100	3
61	winowhy	100	2
	Sum	5384	-

Table 11: Statistics of BIG-Bench sub dataset. We filter the original dataset by retrieving tasks with keywords "multiple choice" and randomly sample up to 100 examples per task. Note, the task in BBH will not be involved in BB-sub.

the rationale and concatenates it with the question to re-predict the answer; (3) output-no correction: Even when the student model identifies rationale errors, the original rationale is used for prediction without any correction. The experimental results indicate that there is no significant performance difference between these three settings, suggesting that the student model is almost incapable of effectively verifying the correctness of the generated rationale. We believe the student model's limited capacity, due to its smaller number of parameters, prevents it from independently verifying the correctness of the rationale, especially in complex reasoning tasks. Additionally, the model may struggle to generalize to different types of reasoning errors.

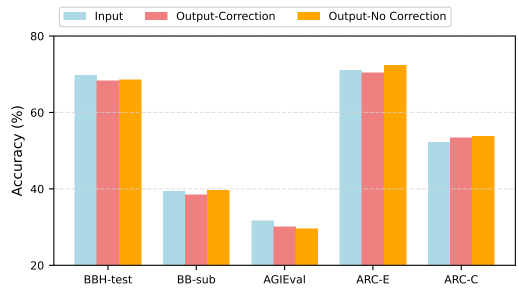

Figure 8: Comparison between using the rationale status string rs_f as input and output in the correction task.

Does the student model have the ability to correct erroneous rationale? We assume that the student model can correct verify erroneous rationales to evaluate its ability to correct them. During evaluation, the student model attempts to correct the rationales corresponding to previously erroneous answer predictions and then concatenates the corrected rationale with the question to re-predict the answer. As shown in Figure 9, The student model's accuracy improves on both the IND and OOD datasets, mainly because 15% to 30% of the previous incorrect predictions are now correct. This suggests that the model can partially correct erroneous rationales, enhancing the final answer accuracy. Although the student model shows some limitations in correcting errors, this finding still reveals the substantial potential of distilling the ability to correct erroneous reasoning into student model.

F Case Study

Table 14 shows that Std-CoT generates incorrect intermediate reasoning steps, leading to an incorrect final answer, indicating that Std-CoT struggles with effective reasoning in complex tasks. In contrast, CoPeD-TL generates a CoT that outperforms the teacher's reasoning. Tables 15 and 16 demonstrate that the intermediate reasoning steps generated by Std-CoT in domain-specific tasks lack causal relationships with the final answers, sug-

system content	`You are a helpful and precise assistant for following the given instruction.`

user content	`[Instruction]{Please correct the wrong rationale by using better reasoning steps.}`
	
	`Task Description:{Task Description}`
	
	`Question: {Question}`
	
	`Answer: {Answer}`
	
	`Wrong rationale: {Wrong rationale}`
	
	`Better Reasoning:`

Table 12: Prompt template for gpt-3.5-turbo for ask the teacher LLM to generate correct rationales.

system content	`You are a helpful and precise assistant for following the given instruction.`

user content	`[Instruction]{Please read the question, rationale, and answer, and simply determine whether the answer can be derived from the rationale. Respond with 'yes' or 'no', without any explanations}`
	
	`Task Description: {Task Description}`
	
	`Question: {Question}`
	
	`Rationale: {Rationale}`
	
	`Answer: {Answer}`

Table 13: Prompt template of evaluator for predicting whether the rationale supports the answer, given the question, rationale, and answer.

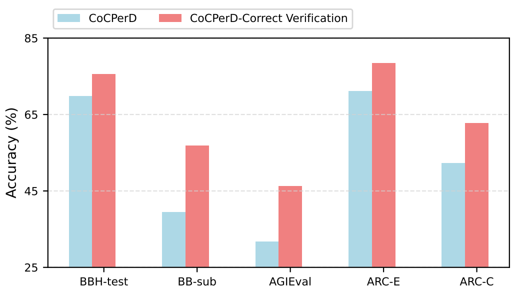

Figure 9: Comparison between CoPeD without correcting rationale errors and correcting rationale errors under the assumption of correct rationale verification.

gesting that Std-CoT causes the student model to capture spurious correlations between the question and the answer. In contrast, the CoT generated by CoPeD-TL effectively supports answer prediction, indicating an enhancement in the student's reasoning ability. Tables 17 and 18 further show that in domain-adapted tasks, Std-CoT almost fails to reason correctly, whereas CoPeD-TL can clearly understand the question and provide logical and coherent reasoning.

Question:	Michael lies. Leda says Michael lies. Delbert says Leda tells the truth. Tamika says Delbert tells the truth. Fidel says Tamika lies. Does Fidel tell the truth?
Teacher	(1) Michael lies. So, we know that Michael lies. (2) Leda says Michael lies. Since we know from (1) that Michael lies, if Leda says Michael lies, then Leda tells the truth. (3) Delbert says Leda tells the truth. Since we know from (2) that Leda tells the truth, if Delbert says Leda tells the truth, then Delbert tells the truth. (4) Tamika says Delbert tells the truth. Since we know from (3) that Delbert tells the truth, if Tamika says Delbert tells the truth, then Tamika tells the truth. (5) Fidel says Tamika lies. Since we know from (4) that Tamika tells the truth, if Fidel says Tamika lies, then Fidel lies. Now, the question asks: Does Fidel tell the truth? We know from (5) that Fidel lies. Therefore, the answer is No. ✓
Std-CoT	According to this question, the answers to the questions are as follows: Michael lies: true. Leda lies: true. Delbert lies: true. Tamika lies: true. Fidel tells the truth: true. We know that Michael lies. Leda lies. Delbert lies. Tamika lies. Fidel tells the truth. To answer this question, we need to evaluate the statement Fidel tells the truth? Yes, Fidel tells the truth. Therefore the answer is Yes. ✗
CoPeD	We first identify the parties in this dialogue as follows: M̈ichael≐ M, L̈eda≐ L, D̈elbert≐ D, T̈amika≐ T, and F̈idel≐ F. The statement M̈ichael liesïs expressed as M = lies. The statement L̈eda says Michael liesïs expressed as L = lies. The statement D̈elbert says Leda tells the truthïs expressed as D = tells-the-truth. The statement T̈amika says Delbert tells the truthïs expressed as T = tells-the-truth. The statement F̈idel says Tamika liesïs expressed as F = lies. Now, the question asks: Does Fidel tell the truth? We know from earlier that F = lies. Therefore, the answer is No. ✓

Table 14: A web of lies case from BIG-Bench Hard.

Question:	Here is a table where the first line is a header and each subsequent line is a penguin: name, age, height (cm), weight (kg) Louis, 7, 50, 11 Bernard, 5, 80, 13 Vincent, 9, 60, 11 Gwen, 8, 70, 15 For example: the age of Louis is 7, the weight of Gwen is 15 kg, the height of Bernard is 80 cm. We now add a penguin to the table: James, 12, 90, 12 And here is a similar table, but listing giraffes: name, age, height (cm), weight (kg) Jody, 5, 430, 620 Gladys, 10, 420, 590 Marian, 2, 310, 410 Donna, 9, 440, 650 How many penguins are more than 5 years old? Options:(A) 1 (B) 2 (C) 3 (D) 4 (E) 5
Teacher	This question focuses on age. We know the following: Louis is 7 years old, Bernard is 5 years old, Vincent is 9 years old, Gwen is 8 years old, and James is 12 years old. The penguins that are more than 5 years old are Louis, Vincent, Gwen, and James. There are 4 penguins more than 5 years old. Therefore, the answer is (D). ✓
Std-CoT	This question focuses on age. We know the following: Louis is 7 years old, Bernard is 5 years old, Vincent is 9 years old, and James is 12 years old. Gwen's age is not mentioned in the table. The penguins that are more than 5 years old are Louis, Bernard, Vincent, and James. There are 4 penguins that are more than 5 years old. Therefore, the answer is (D). ✗
CoPeD	This question focuses on age. We know the following: Louis is 7 years old, Bernard is 5 years old, Vincent is 9 years old, Gwen is 8 years old, and James is 12 years old. According to the table, we can observe that there are 4 penguins that are more than 5 years old: Louis, Vincent, Gwen, and James. Therefore, the answer is (D). ✓

Table 15: A penguins in a table case from BIG-Bench Hard.

Question:	On the desk, you see a fuchsia dog leash and a teal necklace. Is the dog leash turquoise? Options: (A) yes (B) no
Teacher	According to this question, the color of the dog leash is fuchsia, not turquoise. Therefore, the answer is (B) no. ✓
Std-CoT	According to this question, the color of the dog leash is fuchsia. Therefore, the answer is (A) yes. ✗
CoPeD	According to this question, the color of the dog leash is fuchsia. Therefore, the answer is (B) no. ✓

Table 16: A reasoning about colored objects case from BIG-Bench Hard.

Question:	Please answer the following questions about which words certain pronouns refer to. The sun was covered by a thick cloud all morning, but luckily, by the time the picnic started, it was gone. The 'it' refers to the cloud because The cloud covered the sun. The above reasoning is Options: (A) correct (B) incorrect
Teacher	First, we are given the information that the sun was covered by a thick cloud all morning. Then, we are told that by the time the picnic started, it was gone. The pronoun "it" is used to refer to something that was previously mentioned. In this case, the only thing mentioned that could be referred to by "it" is the cloud. Therefore, the answer is (A) correct. ✓
Std-CoT	The answer is not (A) because (A) states that the "it" refers to the picnic, which is not correct based on the context of the story. The "it" refers to the cloud because the cloud covered the sun. Therefore, the answer is (B) ✗
CoPeD	In this question, we are told that "The cloud covered the sun." A pronoun is used in the sentence to refer to the cloud. The sun is not directly mentioned in the sentence, but it is implied that the cloud covered the sun. Therefore, the cloud covered the sun. Therefore, the answer is (A) correct. ✓

Table 17: A winowhy case from BIG-Bench sub.

Question:	Context and question: "But consider your daughters. Only think what an establishment it would be for one of them. Sir Adams and Lady Nelson are determined to go, merely on that account, for in general, you know, they visit no newcomers. Indeed you must go, for it will be impossible for us to visit him if you do not." "You are over-scrupulous, surely. I dare say Mr. Torres will be very glad to see you; and I will send a few lines by you to assure him of my hearty consent to his marrying whichever he chooses of the girls; though I must throw in a good word for my little Helga" "I desire you will do no such thing. Helga is not a bit better than the others; and I am sure she is not half so handsome as Olga, nor half so good-humoured as Freya. But you are always giving her the preference." What is Freya to Olga? Options: (A) Brother (B) Sister (C) Colleague (D) Master (E) Servant (F) Enemy (G) Lover (H) Husband (I) Father-in-law (J) Daughter-in-law (K) Mother-in-law (L) Son (M) Father (N) Wife (O) Grandfather (P) Grandson (Q) Grandmother (R) Son-in-law (S) Sister-in-law (T) Mother (U) Daughter (V) Cousin (W) Granduncle (X) Grandnephew (Y) Brother-in-law
Teacher	First, we are given that Sir Adams and Lady Nelson are determined to go to visit someone. This suggests that the person they are visiting is important or has a high social status. Next, it is mentioned that Mr. Torres is the person they are visiting, and the speaker wants to send a message to him through the person they are addressing. The speaker also mentions that they give a good word for their "little Helga" but the person they are addressing disagrees and says that Helga is not better than the others. Based on this context, we can conclude that Freya, who is mentioned as being better-humored than Helga, is one of the daughters of the person being addressed. Therefore, the answer is (U) Daughter. ✗
Std-CoT	the answer is not (B) or (C), as these are not the correct relationships. the answer is (D), because Freya is the daughter of Sir Adams and Lady Nelson, making her their daughter-in-law. Therefore, the answer is D. ✗
CoPeD	We have (1) "But consider your daughters. Only think what an establishment it would be for one of them." (2) "Sir Adams and Lady Nelson are determined to go, merely on that account, for in general, you know, they visit no newcomers." (3) "You are over-scrupulous, surely. I dare say Mr. Torres will be very glad to see you; and I will send a few lines by you to assure him of my hearty consent to his marrying whichever he chooses of the girls." (4) "You are always giving her the preference." From these clues, we can deduce that Freya is Olga's sister. Therefore, the answer is (B). ✓

Table 18: A implicit-relations case from BIG-Bench sub.