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

**Anonymous ACL submission**


## Abstract

Large language models (LLMs) have demonstrated extraordinary reasoning prowess, but their deployment costs are prohibitively high. Therefore, previous research efforts have endowed Small language models (SLMs) with reasoning abilities by fine-tuning them on the Chain-of-Thought (CoT) data generated by LLMs. However, during the learning process, SLMs may capture spurious correlations between questions and answers, making it difficult to ensure the soundness of the generated rationales and their consistency with the predicted answers. In this work, we propose the Chain-of-Thought Correctness Perception Distillation (CoCPerD) method, which perceives the correctness of the rationale and applies distinct strategies accordingly. Specifically, we collect both correct and erroneous rationales from the teacher and student models. During training, we label the rationales with a status string indicating whether they are correct or erroneous. If the rationale is correct, the student model predicts the answer; if the rationale is erroneous, the student model corrects the erroneous rationale. This encourages the student model to rely on valid reasoning paths for answer prediction and learn from mistakes, thereby enhancing the faithfulness and soundness of the generated rationales. Experiments have shown that CoCPerD is effective on both in-distribution (IND) and out-of-distribution (OOD) benchmark reasoning datasets.


## 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., 2023), the CoT paradigm enables multi-step logical reasoning through explicit intermediate derivations. Although this paradigm enhances complex problem-solving through explicit logical derivations, its computational intensity results in deployment bottlenecks (Shao and Li, 2025). For example, the 175B parameter GPT-3 architecture (Brown et al., 2020) requires substantial computational resources during reasoning, leading to prohibitive operational costs.

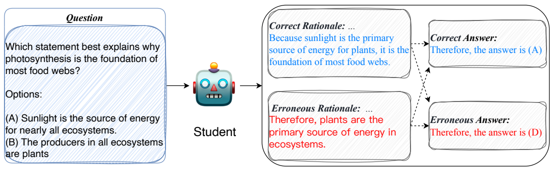


Figure 1: Due to the potential for the student model to capture spurious correlations between the question and the answer during training, there is a possibility that the rationale could be correct while the answer is erroneous, or the answer could be correct while the rationale is erroneous during inference.

Therefore, the current research on knowledge distillation (Magister et al., 2023; Ho et al., 2023; Fu et al., 2023; Zhou and Ai, 2024) 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. Subsequently, these rationales are used to perform supervised fine-tuning on the SLM. Although this paradigm improves the accuracy of SLMs on specific tasks, during the learning process, the SLMs may capture spurious correlations between question and answer, leading to two key limitations. First, these spurious correlations may cause SLMs to overlook the inherent causal logic-based dependencies between rationales and answers, making it impossible to ensure that the generated rationales are consistent with the model's predictions or faithfully justify the decision-making process (Wang et al., 2023a; Feng et al., 2024), as illustrated in Figure 1. Second, these spurious correlations negatively affect the quality of rationale generation during the reasoning process (Dai et al., 2024b), introducing the risk of error propagation. If an error occurs during the intermediate steps of reasoning, subsequent reasoning will be negatively impacted.

To address the above issues, we propose the **C**hain-**o**f-Thought **C**orrectness **Per**ception **D**istillation (CoCPerD), which employs different training strategies based on the correctness of the rationale, allowing the student model (SLM) to learn from mistakes and thereby enhancing the faithfulness and soundness of the generated rationale. Specifically, (1) During the data collection phase, we begin by having the teacher model (LLM) generate both correct and erroneous rationales, allowing the student model to learn from both. Next, we collect the errors made by the student model, with the teacher model providing corresponding corrections. Afterward, we combine the erroneous rationales generated by the student model with the corrections from the teacher model, thereby enriching the teacher model's dataset with both correct and erroneous rationales. This integrated data is subsequently used to retrain the student model. (2) During the training phase, We append a rationale status string as a suffix to indicate the correctness of the rationale, distinguishing between the answer prediction task and the rationale correction task. When the rationale is correct, the student model outputs the corresponding answer based on the question and the correct rationale; when the rationale is erroneous, the student model outputs the corrected rationale. In this way, the student model learns to rely on valid reasoning paths for answer prediction, rather than merely relying on superficial correlations between the question and the answer, thereby enhancing the faithfulness of the generated rationale. Furthermore, the rationale correction task encourages the student model to learn from mistakes, thereby reducing the probability of generating incorrect reasoning steps during inference, which enhances the soundness of the generated rationale.

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

- We propose adopting different training strategies based on the correctness of the rationale, making the student model aware that answer predictions should be grounded in valid rationales, thereby ensuring the consistency between the generated rationale and the predicted answer, and enhancing the faithfulness of the generated rationale.
- We also introduce a task to correct erroneous rationales, allowing the student model to learn from its mistakes, thereby reducing the probability of generating incorrect reasoning steps during inference and improving the soundness of the generated rationale.
- Extensive experiments validate the effectiveness of CoCPerD across both IND and OOD datasets.

## 2 Method

The core idea of our method is to train the student model using different strategies based on the correctness of the rationale. This allows the student model to: 1) become aware of the need to predict answers based on correct rationales, ensuring the faithfulness of the generated rationale; and 2) learn from the task of correcting erroneous rationales, enhancing the soundness of the generated rationale. An overview of our method is shown 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. These rationales are collected 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 provide prompt templates to guild the teacher generate correct and erroneous rationales with similar reasoning paths but different conclusions. Eventually, we construct the dataset $\mathcal{D}^{\text{T}}_{\text{train}} = \{(q_i, r_i^+, r_i^-, a_i)\}_{i=1}^{n}$ 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.

### 2.2 Collection Rationales using Student

The Std-CoT method may cause the student model to capture spurious correlations between the ques-

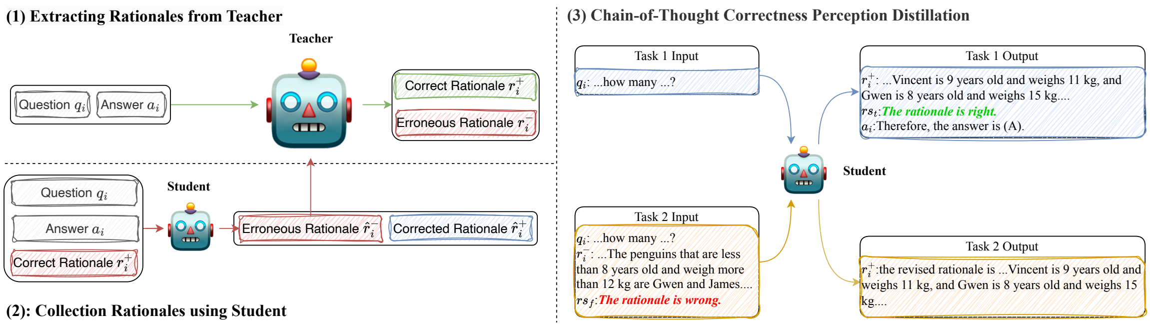


Figure 2: Overview of **Chain-of-Thought Correctness Perception Distillation(CoCPerD)**. We use teacher and student models to generate correct and erroneous rationales for the entire training set. Then, we adopt a multi-task learning framework to leverage these rationales, where one task is trained to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

tion and the answer during the training phase,, which could negatively impact the quality of the rationale generated during inference. (Dai et al., 2024b). To mitigate this, we have the teacher model correct the student model's mistakes, allowing the student model to learn from them, which helps: (1) improve rationale quality and (2) encourage deeper reflection on errors.

**Fine-tuning SLM with Correct Rationales** We fine-tune the student model $\pi$ using correct rationales $r_i^+$ and answers $a_i$ to obtain the fine-tuned model $\pi_{sft}$. The training objective is to minimize the negative log-likelihood of the sequence of rationale $r_i^+$ and answer $a_i$:

$$\mathcal{L}_{\mathrm{Std-CoT}} = \mathbb{E}_{(q,r^+)\sim\mathcal{D}^{\mathrm{T}}_{\mathrm{train}}}[\ell(q, r^+ \oplus a)], \quad (1)$$

where $\ell$ signifies the negative log-likelihood loss function, expressed as:

$$\ell(x, y) = -\sum_{y_t \in y} \log P\left(y_t \mid x, y_{<t}\right) \quad (2)$$

**Collect the Erroneous Rationale of Student** To evaluate the limitations of the student model $\pi_{sft}$, we test it on the training set $\mathcal{D}_{\mathrm{train}}$. The student model generates rationales $\hat{r}$ and answers $\hat{a}$. If the predicted answer $\hat{a}_i$ differs from the true answer $a_i$, the rationale is considered erroneous. We collect erroneous samples as follows:

$$\mathcal{D}^{\mathrm{S}}_{\mathrm{neg}} = \{(q_i, \hat{r}_i, a_i) \mid \hat{a}_i \neq a_i, (q_i, a_i) \in \mathcal{D}_{\mathrm{train}}\} \quad (3)$$

**Correct the Erroneous Rationale of Student** To correct the erroneous rationales of the student model $\pi_{sft}$, we first construct a prompt[1] $p_c$, based on the student's erroneous rationale $\hat{r}_i^-$ and the correct answer $a_i$, to guide the LLM-based correction process. The erroneous rationales are then corrected as follows:

$$\hat{r}_i^+ = f_{\mathrm{correct}}\left(p_c, q_i, \hat{r}_i^-, a_i\right), \quad (4)$$

where $f_{\mathrm{correct}}$ is the correction function that takes the erroneous rationale and corrects it.

We then pair the student model's erroneous rationales with the teacher model's corrections to form the training dataset $\mathcal{D}^{\mathrm{S}}_{\mathrm{train}} = \{(q_i, \hat{r}_i^+, \hat{r}_i^-, a_i)\}_{i=1}^{m}$. Finally, we combine $\mathcal{D}^{\mathrm{T}}_{\mathrm{train}}$ and $\mathcal{D}^{\mathrm{S}}_{\mathrm{train}}$ to create the final training dataset $\mathcal{D}^{\mathrm{M}}_{\mathrm{train}} = \{(q_i, \hat{r}_i^+, \hat{r}_i^-, a_i)\}_{i=1}^{n+m}$, ready for the reasoning self-validation distillation.

### 2.3 Chain-of-Thought Correctness Perception Distillation

To enable the student model to learn how to adjust its prediction strategy based on the correctness of the rationale, we propose a multi-task learning framework consisting of two training tasks: answer prediction and rationale correction. When the rationale is correct, the student model utilizes it to predict the answer; when the rationale is erroneous, the student model generates a corrected rationale. During training, we append rationale status strings $rs_{\mathrm{t}}$ and $rs_{\mathrm{f}}$ as suffixes to the rationale to distinguish between the task types. As illustrated in Figure 3, reasoning with correctness perception requires the student to predict answers based not only on the question but also on the correct rationale, thereby teaching the student to reason faithfully, that is, to provide reasonable answers grounded in the 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

[1]The prompt for correcting the erroneous rationale is provided in Appendix B.1.

$rs_\text{t}$ = **"the rationale is right"** and the answer $a$. The loss function for the answer prediction task is formulated as follows:

$$\mathcal{L}_\text{A} = \mathbb{E}_{(q,r^+,a)\sim\mathcal{D}^\text{M}_\texttt{train}} \left[\ell(q, r^+ \oplus rs_\text{t} \oplus a)\right] \quad (5)$$

In the rationale correction task, we concatenate the question $q$, the erroneous rationale $r^-$ and the rationale status string $rs_\text{f}$ = **"the rationale is wrong"** as the student model input, where the output label is the correct rationale $r^+$. This task design aims to enable the student model to learn to correct erroneous rationales, thereby implicitly enhancing 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:

$$\mathcal{L}_\text{R} = \mathbb{E}_{(q,r^+,r^-)\sim\mathcal{D}^\text{M}_\texttt{train}} \left[\ell(q \oplus r^- \oplus rs_\text{f}, r^+)\right] \quad (6)$$

The final optimization process integrates the loss $\mathcal{L}_\text{A}$ from the answer prediction task and the loss $\mathcal{L}_\text{R}$ from the rationale correction task. Consequently, the combined learning loss from Equation 5 and Equation 6 is formulated as follows:

$$\mathcal{L}_\text{CoCPerD} = (1-\alpha)\mathcal{L}_\text{A} + \alpha\mathcal{L}_\text{R}\,, \quad (7)$$

where $\alpha$ is the hyperparameter used to weight the losses between the two learning tasks.

It is important to note that under the current training setup, the student model cannot predict whether the generated rationale is correct during inference. This is because, during training, we treat the rationale status string $rs_\text{f}$ as part of the student model's input in the rationale correction task, rather than as the target output. When the rationale status string $rs_\text{f}$ 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 §3.6, we further discuss the specific impact on student model performance when the rationale status string $rs_\text{f}$ is used as the student model's input and output in the rationale correction task, respectively. Additionally, we provide a detailed explanation of the reasoning process of the student model under these two settings in the Appendix C.

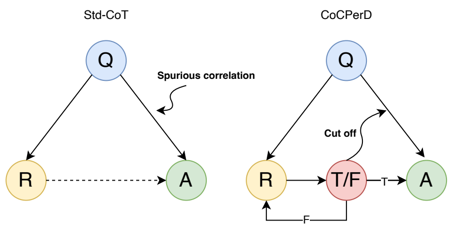


Figure 3: CoCPerD adopts different strategies based on the correctness of the rationale, cutting off the spurious correlation between the question and the answer.

## 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 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** In our main experiment, we use the widely adopted open-source language model

| 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** | | | | | | | | |
| Zero-shot♠ | × | × | 14.8 | 15.5 | 6.9 | 18.2 | 13.9 | 13.9 |
| Zero-shot-CoT♠ | × | ✓ | 10.6 | 7.7 | 7.1 | 18.4 | 14.8 | 11.7 |
| Answer-SFT | × | × | 51.2 | 33.6 | 30.8 | 72.1 | 53.5 | 48.2 |
| 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 |
| EDIT (Dai et al., 2024a) | ✓ | ✓ | 61.5 | 32.3 | 26.7 | 63.9 | 51.0 | 47.1 |
| CasCoD (Dai et al., 2024b) | ✓ | ✓ | 60.2 | 37.2 | 28.6 | 71.1 | 52.4 | 49.9 |
| CoCPerD w/ $\mathcal{D}^{\mathrm{T}}_{\mathrm{train}}$ | ✓ | ✓ | 63.1 | 38.3 | 30.2 | **72.6** | **55.1** | 51.8 |
| CoCPerD w/ $\mathcal{D}^{\mathrm{M}}_{\mathrm{train}}$ | ✓ | ✓ | **64.3** | **39.6** | **31.4** | 71.9 | 54.2 | **52.3** |

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**.

LLaMA2-7B (Touvron et al., 2023) as the student model. 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 $\alpha$ 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 A.1.

**Baselines** We compare our method with the following baselines: (1) **Teacher & Vanilla Student** in Zero-shot(Radford et al., 2019), 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) **SCOTT** (Wang et al., 2023a) that enhances the reasoning consistency of the student model by introducing additional counterfactual data. (5) **EDIT** (Dai et al., 2024a) uses prompts to generate dual CoTs data with similar reasoning paths but different conclusions, then applies the minimum edit distance algorithm to locate and optimize key reasoning steps. (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, CoCPerD achieves state-of-the-art performance on both IND and OOD benchmarks. Specifically, LLaMA2-7B with CoCPerD attains an average accuracy of 52.3% across all tasks, outperforming the strongest baseline (CasCoD) by 2.4%. Notably, CoCPerD exhibits remarkable generalizability in OOD scenarios: On BB-sub, AGIEval, ARC-E, ARC-C, it surpasses CasCoD by 2.4%, 2.8%, 1.5%, and 2.7%, respectively. This indicates that CoCPerD enables the student model to realize the need to predict answers based on correct rationales, allowing the student model to benefit from the generated rationales during answer prediction, thereby enhancing the faithfulness of reasoning.

In addition, CoCPerD w/ $\mathcal{D}^{\mathrm{M}}_{\mathrm{train}}$ outperforms CoCPerD w/ $\mathcal{D}^{\mathrm{T}}_{\mathrm{train}}$ by 1.2% on the IND and also achieves competitive results on the OOD. This indicates that allowing the student model to learn from its own mistakes effectively improves the quality of the generated rationales, ultimately enhancing overall performance.

### 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 evaluate whether the rationale provided by the student model supports its prediction (i.e., faithfulness) and 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[2] $p_e$ to guide LLM-based scoring. We define faithfulness and soundness as follows:

$$\text{Faithfulness} = \mathbb{E}\big[f_{\text{eval}}(p_e, q_i, \hat{r}_i, \hat{a}_i)\big], \quad (8)$$

$$\text{Soundness} = \mathbb{E}\big[f_{\text{eval}}(p_e, q_i, \hat{r}_i, a_i)\big], \quad (9)$$

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 |
| CoCPerD(ours) | 82.4 | 76.3 | 79.4 | 71.3 | 63.3 | 67.3 |

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 CoCPerD is more consistent with the answer (including both the predicted and the ground-truth answers). This indicates that CoCPerD ensures the faithfulness and soundness of the rationale generated during the reasoning process by adopting different strategies based on the correctness of the rationale.

### 3.5 Ablation Study

**Model Size** We conducted model distillation on TinyLLaMA-1.1B[3] (Zhang et al., 2024), LLaMA2-7B, and LLaMA2-13B, and compared it with standard CoTs distillation (Std-CoT), multi-task distillation (MT-CoT), and cascade distillation (CasCod). As shown in Figure 4, we observed that CoCPerD achieves competitive results across student models of various sizes compared to baseline methods, and performs exceptionally well on both IND and OOD datasets. Particularly on the IND dataset, the 1.1B model with CoCPerD reaches 96.5% of the teacher model's performance, demonstrating the significant advantages of CoCPerD in low-resource scenarios. Furthermore, across different model sizes, CoCPerD achieves competitive performance on the OOD dataset compared to the baseline models.

**Data Size** CoCPerD demonstrates significant improvements over baseline methods on both IND and OOD datasets, while utilizing considerably less training data. As shown in Figure 5, CoCPerD achieves a 13.6% 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 instance, on the BB-sub dataset, CoCPerD surpasses CasCoD—trained with the full dataset—by using just 12.5% of the full BBH-train data. On other OOD datasets, CoCPerD also achieves excellent performance. These results clearly demonstrate the effectiveness of CoCPerD in low-resource settings, highlighting its ability to enhance the performance of CoTs both IND and OOD with significantly less training data.

### 3.6 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 6, 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 CoCPerD 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.

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

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

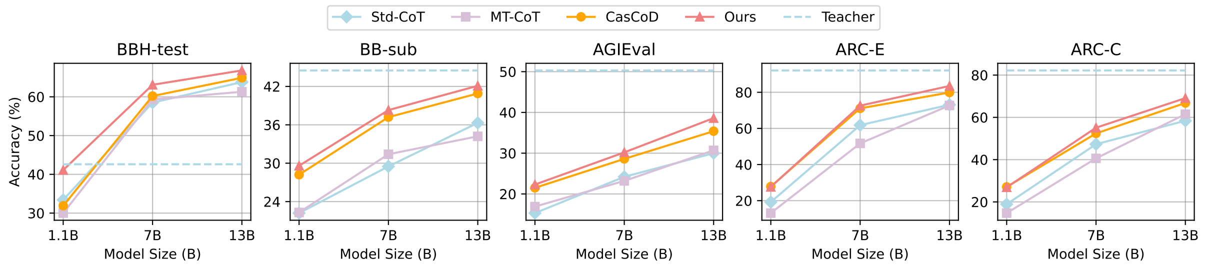


Figure 4: Ablation study on model size for IND and four OOD datasets. The dotted line indicates the performance of the teacher LLM under the Zero-shot-CoT setting.

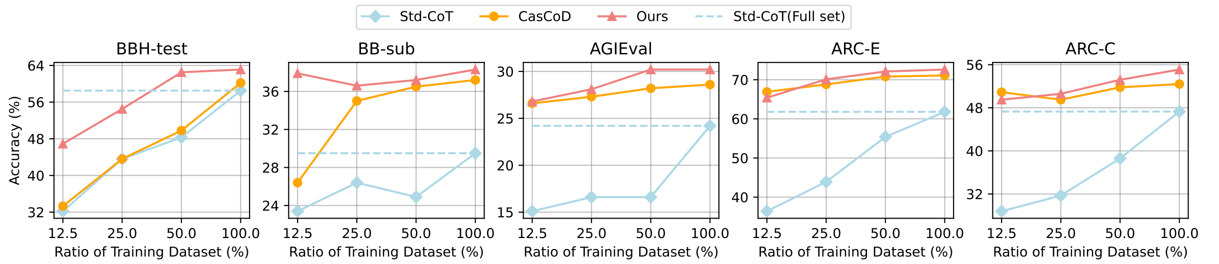


Figure 5: Ablation study on training data size for IND and four OOD datasets. The dotted line indicates the performance of fine-tuning the student models by Std-CoTs distillation using the full set (100% of) BBH-train dataset.

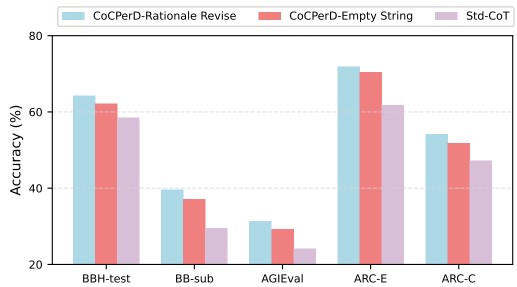


Figure 6: Compare training CoCPerD with different target outputs when the rationale is erroneous.

**Whether the student model can effectively verify the correctness of the rationale?** We explore the impact of using the rationale status string $rs_{\mathrm{f}}$ as input and output on the model's performance in the rationale correction task on the IND and OOD datasets. As shown in Figure 7, the experiment includes the following three settings: (1) input: When the rationale status string $rs_{\mathrm{f}}$ 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 $rs_{\mathrm{f}}$ is used as output, the student model, after identifying rationale errors, corrects 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.

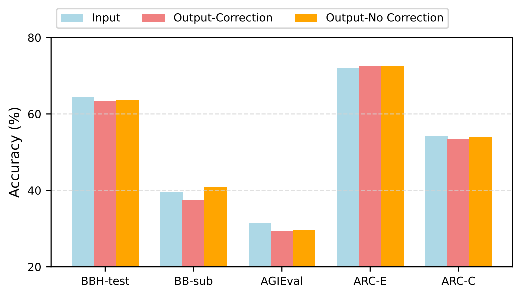


Figure 7: Comparison between using the rationale status string $rs_{\mathrm{f}}$ as input and output in the correction task.

**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 8, The student model's accuracy improves on both the IND and OOD datasets, mainly because 20% to 25% 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.

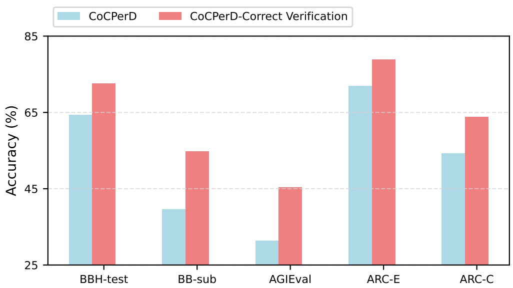


Figure 8: Comparison between CoCPerD without correcting rationale errors and correcting rationale errors under the assumption of correct rationale verification.

### 3.7 Case Study

Due to page limitations, we provide a systematic case study in Appendix D to illustrate the improvement in CoT faithfulness and soundness.

## 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. Liao et al. (2024) propose leveraging symbolic knowledge bases (KB) to enhance the SLM's performance on complex reasoning tasks.

**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.

## 5 Conclusion

In this study, we propose a Chain-of-Thought Correctness Perception Distillation method (CoCPerD). This method utilizes both teacher and student models to collect correct and erroneous rationales, and re-trains the student model to adopt different strategies based on the correctness of the rationale. This effectively mitigates the spurious correlation between questions and answers. By enabling the student model to predict answers based on correct rationales, it enhances the faithfulness of reasoning, while the rationale correction task implicity improves reasoning quality. Experiments demonstrate that CoCPerD significantly outperforms baseline methods on both IND and OOD datasets.

## 6 Limitations

In our study, we explore enabling the student model to verify the correctness of the generated rationale during inference and attempt to correct it when the rationale is identified as erroneous. However, the student model is unable to effectively verify whether the rationale derived from reasoning is correct. Even assuming the model can accurately verify the correctness of the rationale generated during inference, its ability to recover from errors remains limited. This is because verifying the correctness of the rationale and correcting errors is a more complex reasoning task, particularly challenging for SLMs.

Moreover, when collecting correct and erroneous rationales, we determine the correctness of the rationale solely based on whether the model's predicted answer is correct. However, both LLMs and SLMs may exhibit spurious correlations between the question and the answer. As a result, it is possible that a rationale could be correct while the answer is erroneous, or the answer could be correct while the rationale is erroneous, leading to noise in the collected datasets of correct and erroneous samples. Currently, our work does not define a method for effectively filtering out such noisy data. We hope that our research can inspire further exploration in this area and leave this challenge for future investigations.

## References

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. 2023. Mcc-kd: Multi-cot consistent knowledge distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6805–6820.

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, Miami, Florida, USA. Association for Computational Linguistics.

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. *arXiv preprint arXiv:2203.15556*.

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.

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*.

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*.

Huanxuan Liao, Shizhu He, Yao Xu, Yuanzhe Zhang, Kang Liu, and Jun Zhao. 2024. Neural-symbolic collaborative distillation: Advancing small language models for complex reasoning tasks. *arXiv preprint arXiv:2409.13203*.

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).

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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.

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.

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.

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*.

# A Experimantal Settings

## A.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 CoCPerD 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 3 and Table 4, 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, 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.

| **Hyperparameter** | **TinyLLaMA-1.1B** | **LLaMA2-7B** | **LLaMA2-13B** |
|---|---|---|---|
| gradient accumulation steps | 2 | 2 | 2 |
| per device batch size | 2 | 2 | 2 |
| learning rate | 4e-4 | 3e-4 | 1e-4 |
| epoches | 20 | 15 | 15 |
| max length | 1024 | 1024 | 1024 |
| $\beta$ of AdamW | (0.9,0.999) | (0.9,0.999) | (0.9,0.999) |
| $\epsilon$ of AdamW | 1e-8 | 1e-8 | 1e-8 |
| $\gamma$ of Scheduler | 0.95 | 0.95 | 0.95 |
| weight decay | 0 | 0 | 0 |
| warmup ratio | 0 | 0 | 0 |
| rank of LoRA | 64 | 64 | 64 |
| $\alpha$ of LoRA | 32 | 32 | 32 |
| target modules | q_proj, v_proj | q_proj, v_proj | q_proj, v_proj |
| drop out of LoRA | 0.05 | 0.05 | 0.05 |

Table 3: 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 4: Generation configs of students and teachers.

## A.2 Dataset Statistics

Table 5, Table 6, Table 7 and Table 8 show the data statistics of AGIEval, ARC, BIG-Bench Hard (BBH) and BIG-Bench Sub (BB-sub), respectively.

| **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 5: Statistics of AGIEval dataset.

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

Table 6: Statistics of ARC test dataset.

| 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 7: Statistics of BIG-Bench Hard dataset.

| 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 8: 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.

# B Prompts

## B.1 Prompts of Correct the Erroneous Rationale for ChatGPT

We use the prompt template shown in Table 9 to call the ChatGPT API to correct the erroneous rationale of student model for the BBH-train datasets.

## B.2 Prompts of Evaluator

We use the prompt templates shown in Table 10 to call the ChatGPT and GPT-4 APIs, predicting whether the rationale supports the answer.

# C Inference Process

Figure 9 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 $rs_{\mathrm{f}}$ 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 $rs_{\mathrm{f}}$ 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 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

| 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.}<br><br>Task Description:{Task Description}<br><br>Question: {Question}<br><br>Answer: {Answer}<br><br>Wrong rationale: {Wrong rationale}<br><br>Better Reasoning: |

Table 9: 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}<br><br>Task Description: {Task Description}<br><br>Question: {Question}<br><br>Rationale: {Rationale}<br><br>Answer: {Answer} |

Table 10: Prompt template of evaluator for predicting whether the rationale supports the answer, given the question, rationale, and answer.

question and the verification string $rs_{\mathrm{t}}$ as input to predict the answer.

## D Case Study

Table 11 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, CoCPerD generates a CoT that outperforms the teacher's reasoning. Tables 12 and 13 demonstrate that the intermediate reasoning steps generated by Std-CoT in domain-specific tasks lack causal relationships with the final answers, suggesting that Std-CoT causes the student model to capture spurious correlations between the question and the answer. In contrast, the CoT generated by CoCPerD effectively supports answer prediction, indicating an enhancement in the student's reasoning ability. Tables 14 and 15 further show that in domain-adapted tasks, Std-CoT almost fails to reason correctly, whereas CoCPerD can clearly understand the question and provide logical and coherent reasoning.

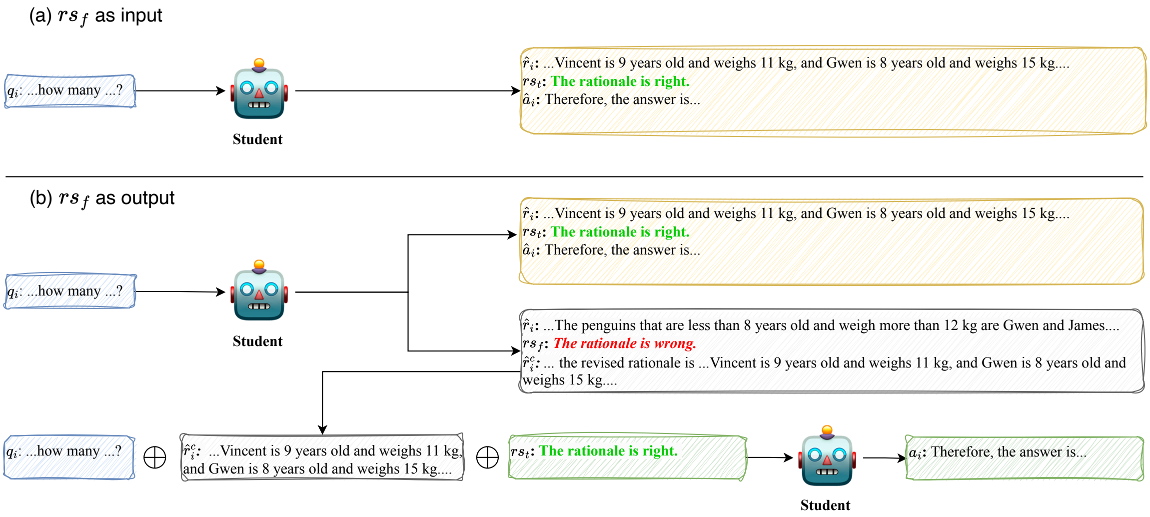


Figure 9: Comparison of student model inference processes under different training strategies.

| **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.<br>(2) Leda says Michael lies. Since we know from (1) that Michael lies, if Leda says Michael lies, then Leda tells the truth.<br>(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.<br>(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.<br>(5) Fidel says Tamika lies. Since we know from (4) that Tamika tells the truth, if Fidel says Tamika lies, then Fidel lies.<br>Now, the question asks: Does Fidel tell the truth? We know from (5) that Fidel lies.<br>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.<br>We know that Michael lies. Leda lies. Delbert lies. Tamika lies. Fidel tells the truth.<br>To answer this question, we need to evaluate the statement Fidel tells the truth?<br>Yes, Fidel tells the truth.<br>Therefore the answer is Yes. ✗ |
| **CoCPerD** | We first identify the parties in this dialogue as follows: M̈ichael≐ M, L̇eda≐ L, D̈elbert≐ D, T̈amika≐ T, and F̈idel≐ F.<br>The statement M̈ichael liesïs expressed as M = lies. The statement L̈eda says Michael liesïs expressed as L = lies.<br>The statement D̈elbert says Leda tells the truthïs expressed as D = tells-the-truth.<br>The statement T̈amika says Delbert tells the truthïs expressed as T = tells-the-truth.<br>The statement F̈idel says Tamika liesïs expressed as F = lies.<br>Now, the question asks: Does Fidel tell the truth? We know from earlier that F = lies.<br>Therefore, the answer is No. ✓ |

Table 11: 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?<br>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.<br><br>The penguins that are more than 5 years old are Louis, Vincent, Gwen, and James.<br><br>There are 4 penguins more than 5 years old.<br><br>Therefore, the answer is (D). ✓ |
| **Std-CoT** | This question focuses on age. We know the following:<br><br>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.<br><br>The penguins that are more than 5 years old are Louis, Bernard, Vincent, and James.<br><br>There are 4 penguins that are more than 5 years old.<br><br>Therefore, the answer is (D). ✗ |
| **CoCPerD** | This question focuses on age. We know the following:<br><br>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.<br><br>According to the table, we can observe that there are 4 penguins that are more than 5 years old: Louis, Vincent, Gwen, and James.<br><br>Therefore, the answer is (D). ✓ |

Table 12: 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?<br>Options: (A) yes (B) no |
| **Teacher** | According to this question, the color of the dog leash is fuchsia, not turquoise.<br><br>Therefore, the answer is (B) no. ✓ |
| **Std-CoT** | According to this question, the color of the dog leash is fuchsia.<br><br>Therefore, the answer is (A) yes. ✗ |
| **CoCPerD** | According to this question, the color of the dog leash is fuchsia.<br><br>Therefore, the answer is (B) no. ✓ |

Table 13: A reasoning about colored objects case from BIG-Bench Hard.

| | |
|---|---|
| **Question:** | Please answer the following questions about which words certain pronouns refer to.<br>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.<br>The above reasoning is<br>Options: (A) correct (B) incorrect |
| **Teacher** | First, we are given the information that the sun was covered by a thick cloud all morning.<br><br>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.<br><br>In this case, the only thing mentioned that could be referred to by "it" is the cloud.<br><br>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.<br><br>The "it" refers to the cloud because the cloud covered the sun.<br><br>Therefore, the answer is (B) ✗ |
| **CoCPerD** | In this question, we are told that "The cloud covered the sun." A pronoun is used in the sentence to refer to the cloud.<br><br>The sun is not directly mentioned in the sentence, but it is implied that the cloud covered the sun.<br><br>Therefore, the cloud covered the sun.<br><br>Therefore, the answer is (A) correct. ✓ |

Table 14: 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?<br>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.<br><br>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.<br><br>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.<br><br>Therefore, the answer is (U) Daughter. ✗ |
| **Std-CoT** | the answer is not (B) or (C), as these are not the correct relationships.<br><br>the answer is (D), because Freya is the daughter of Sir Adams and Lady Nelson, making her their daughter-in-law.<br><br>Therefore, the answer is D. ✗ |
| **CoCPerD** | We have (1) "But consider your daughters. Only think what an establishment it would be for one of them."<br><br>(2) "Sir Adams and Lady Nelson are determined to go, merely on that account, for in general, you know, they visit no newcomers."<br><br>(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."<br><br>(4) "You are always giving her the preference."<br><br>From these clues, we can deduce that Freya is Olga's sister.<br><br>Therefore, the answer is (B). ✓ |

Table 15: A implicit-relations case from BIG-Bench sub.