# Towards Efficient CoT Distillation: Self-Guided Rationale Selector for Better Performance with Fewer Rationales

**Jianzhi Yan[1,2], Le Liu[1,2], Youcheng Pan[2]*, Shiwei Chen[1,2]**
**Yang Xiang[2,3]*, Buzhou Tang[1,2]***
[1]Harbin Institute of Technology, Shenzhen, China
[2]Pengcheng Laboratory, Shenzhen, China
[3]Shaoguan Research Institute of Data Industry, China
{yanjzh, liul07, panych, chenshw ,xiangy}@pcl.ac.cn
tangbuzhou@gmail.com

## Abstract

Chain-of-thought (CoT) distillation aims to enhance small language models' (SLMs) reasoning by transferring multi-step reasoning capability from the larger teacher models. However, existing work underestimates rationale quality, focusing primarily on data quantity, which may transfer noisy or incorrect information to the student model. To address the above issues, we proposed **M**odel-**O**riented **R**ationale **S**election **D**istillation (MoRSD), which can discern and select high quality rationales for distillation to improve performance further. We further propose a Rationale Difficulty (RD) metric to measure the ability of the student model to generate the correct answer under a given rationale. Compared to the baseline, we achieved 4.6% average improvement on seven datasets over three tasks, using fewer rationales by controlling their accuracy, diversity, and difficulty. Our results reveal that a small portion of the high quality rationales can enhance the reasoning ability of student models than the entire dataset. Our method promises to be a possible solution for efficient CoT distillation. Our code will be released in https://github.com/Leon221220/MoRSD.

## 1 Introduction

Large language models (LLMs) such as LLaMA, GPT-4, Gemini, DeepSeek-V3, and DeepSeek-R1, have achieved remarkable performance in various reasoning tasks by instructing them to *think step-by-step* (Touvron et al., 2023; OpenAI et al., 2024; Zhang et al., 2024a; DeepSeek-AI et al., 2024, 2025; Brown et al., 2020; Sun et al., 2021). Engaging in reasoning through logically coherent steps has substantially enhanced performance in tasks such as mathematical problem solving and question answering. These intermediate reasoning steps are referred to as *rationale* (Wei et al., 2023).

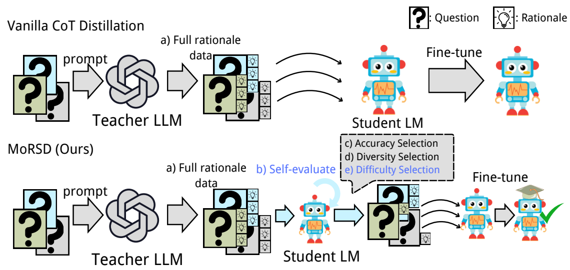


Figure 1: **Vanilla CoT Distillation and MoRSD.** Different from previous studies that mostly use a), c), and d), we propose b) and e) to select effective data for specific student models to improve performance further.

To achieve emergent reasoning abilities, LLMs require large-scale parameters, making SLMs inherently limited (Wei et al., 2023; Kojima et al., 2023; Fu et al., 2023). CoT distillation has become a key technique for enhancing SLM reasoning by transferring rationales from stronger teachers (Wang et al., 2023b; Li et al., 2023), showing strong results on arithmetic and symbolic tasks (Ho et al., 2023; Hsieh et al., 2023; Ying et al., 2024; Kim et al., 2024). Beyond basic distillation, recent works explore consistency enforcement (Chen et al., 2023), cross-task supervision (Li et al., 2024), and tailored strategies (Zhang et al., 2024b). Mentor-KD (Lee et al., 2024) introduces intermediate models for better supervision, MCC-KD promotes consistent yet diverse reasoning (Chen et al., 2023), while Lion (Kim et al., 2024) and TA-in-the-Loop (Zhang et al., 2024b) use adversarial and auxiliary guidance, respectively.

However, these approaches often require additional models, discard useful failures, or introduce iterative overhead—resulting in high computational costs and limited flexibility. And many works still rely on enlarging the rationale set (increasing from 1 to 8 per instance (Ho et al., 2023)) to improve performance, while **overlooking rationale quality**. Such data scaling ignores variance

*Corresponding authors.



*Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 7818–7835
November 4-9, 2025 

in correctness and diversity, risking the distillation of noisy signals. Furthermore, most approaches **neglect the specificity of student models**, failing to adapt to their strengths or limitations. These limitations motivate our focus: how to select a small set of high-quality, student-aware rationales for efficient and effective distillation.

To overcome these limitations, we propose **MoRSD**, a simple but effective method that enables student models to customize their distillation data autonomously. As presented in Figure 1, MoRSD consists of four stages: 1) rationale generation, 2) self-evaluation, 3) rationale selection and 4) distillation. The rationale generation stage prompts the teacher LLM to generate the rationale dataset. In the self-evaluation stage, we calculate rationale difficulty (RD) to measure the contribution of a given rationale to distillation. Specifically, RD measures the student's ability to generate the correct answer given a question and rationale. Those with smaller RD are considered more beneficial to generate the corresponding answer.

Then, we first apply model-agnostic accuracy selection and diversity selection to the rationale dataset. Accuracy selection adjusts the proportion of correct rationales in the dataset to achieve the given accuracy threshold, diversity selection involves pairwise Jaccard similarity to eliminate similar rationale in the dataset. Finally, we use difficulty selection to select the rationales with smaller RD. Since difficulty selection uses perplexity-based RD, a model-specific metric, it enables the student model to customize its distillation data during the difficulty selection. Through these stages, we obtain a small amount of high-quality rationale data to improve distillation performance for specific student models. In summary, our contributions are three-fold:

1. We propose **MoRSD**, a simple and effective method that performs better with fewer rationales. Prove that using a small portion of the dataset can outperform using the entire dataset in enhancing the reasoning ability of student models.

2. We propose a model-specific metric, rationale difficulty, to measure rationale contribution for distillation, enabling student models to customize data based on their training requirements.

3. We conducted extensive experiments on seven datasets covering three distinct tasks. The results demonstrate that our method consistently outperforms the baselines, achieving an average accuracy improvement of 4.6%.

## 2 Related work

### 2.1 Chain-of-thought (CoT) Distillation

Chain-of-thought prompting delivers strong performance but typically benefits from large models with many parameters, resulting in high computational costs and limited deployment (Hoffmann et al., 2022; Chowdhery et al., 2022). Ho et al. (2023) first introduced fine-tune-CoT, a method that transfers the multi-step reasoning ability of LLMs to smaller models through fine-tuning. Some approaches use in-context learning to implicitly transfer knowledge (Rajani et al., 2019; Wang et al., 2023a), while others treat rationale generation as a multi-task fine-tuning objective (Hsieh et al., 2023). Furthermore, Li et al. (2024) distill the rationale into multiple experts in low-rank adaptation (LoRA), decoupling CoT reasoning from the student model. Zhang et al. (2024b) enhances knowledge transfer through active learning and explanation-guided sample selection. Some researchers identify influential tokens using gradient attribution techniques such as saliency maps to guide the student model (Ballout et al., 2024). Recently, a study found that only a small fraction (4.7%) of CoT steps are critical for performance (Dai et al., 2024), which closely matches our findings. Busbridge et al. (2025) introduce a distillation scaling law to optimize compute allocation between teacher and student models, providing efficient distillation strategies that outperform supervised pretraining in certain cases.

### 2.2 Data Efficiency in Language Models

Data efficiency means that the model achieves high performance with a smaller amount of training data, maximizing the value derived from limited data. Yang et al. (2024) shows that with only 1,000 carefully selected prompts and responses, models can learn to follow specific formats and generalize effectively to new tasks. Chen et al. (2024) used GPT-3.5 to score data difficulty, and Mekala et al. (2024) proposed Learning Percentage (LP) for difficulty assessment, both reduced data needs for instruction tuning. LIMA achieves strong performance with few examples, generalizing well to unseen tasks and requiring minimal instruction tuning (Zhou et al., 2023). Yue et al. (2024) uses a multi-round distillation framework with an oracle LLM to select challenging instructions for

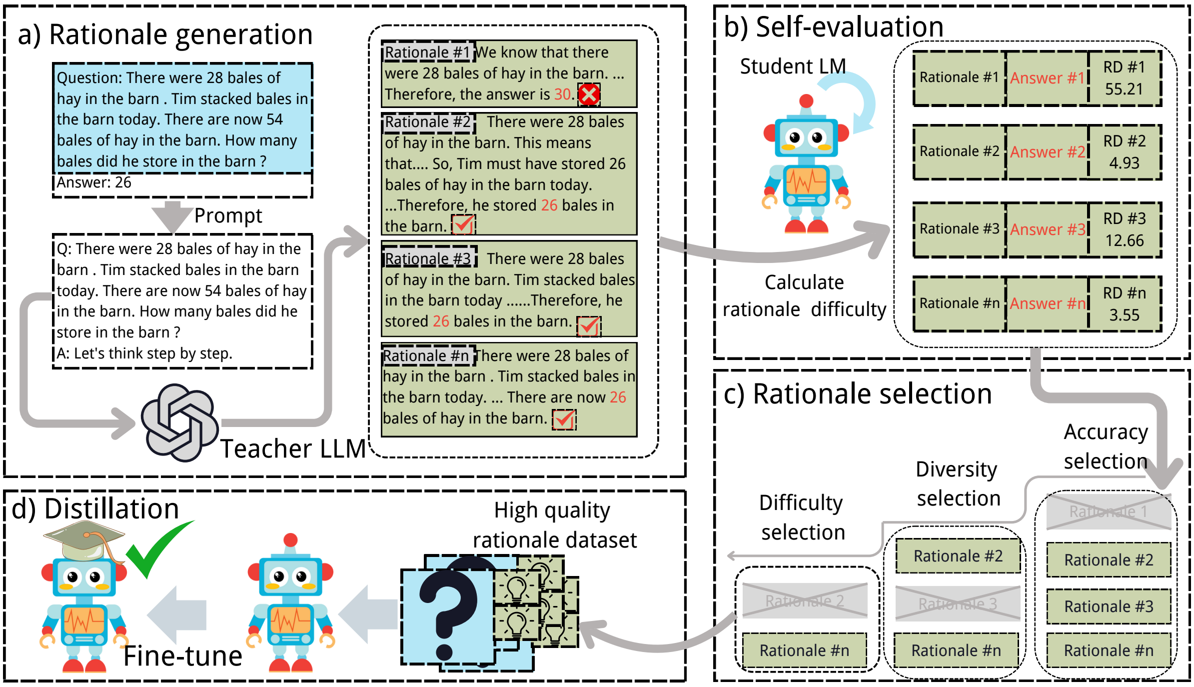


Figure 2: **Detailed overview of MoRSD**. MoRSD comprises four stages: **a) Rationale generation** prompts a teacher model to generate all the data required for the rationale selection stage (blue). **b) Self-evaluation**, which uses the rationale difficulty (RD) to evaluate all generated rationales. Those rationales with smaller RD are considered helpful for distillation. **c) Rationale selection**, which constructs the final dataset for distillation by controlling the original dataset's accuracy, diversity, and difficulty. **d) Distillation**, which fine-tunes the student model using the constructed dataset.

student models, reducing the need for extensive training samples. Recently, Ye et al. (2025) proposed the "Less is More Reasoning Hypothesis" (LIMO), demonstrating that complex reasoning can be induced with few examples when the base model has pre-trained domain knowledge. Muennighoff et al. (2025) introduced a test-time scaling approach using a curated dataset (s1K) and budget forcing, enabling the Qwen2.5-32B-Instruct model to outperform OpenAI's o1-preview (OpenAI, 2024) on math reasoning tasks by 27% with controlled test-time compute.

## 3 Method

### 3.1 Problem definition

CoT distillation first requires prompting the teacher model to generate rationales related to the training data. Let $\mathcal{D} = \{(q_1, a_1), (q_2, a_2), \ldots, (q_N, a_N)\}$ denote the complete dataset, where each $(q_i, a_i)$ represents a question-answer pair and the label is available. Then, a teacher model $\mathcal{T}\left(\theta^{\mathcal{T}}\right)$ (the parameter $\theta^{\mathcal{T}}$ is inaccessible) is prompted to generates $m$ distinct rationale $\{\hat{r}_i^1, \hat{r}_i^2, \ldots, \hat{r}_i^m\}$ where each $\hat{r}_i^j$ represents a separate rationale for the question $q_i$. The complete dataset with these rationales is denoted as:

$$\mathcal{D}_{\text{full}} = \left\{q_i, \{(\hat{r}_i^1, \hat{a}_i^1), (\hat{r}_i^2, \hat{a}_i^2), \ldots, (\hat{r}_i^j, \hat{a}_i^j)\}\right\} \tag{1}$$

Where $i = 1, 2, \ldots, N, j = 1, 2, \ldots, M$. The performance of the student model $\mathcal{S}$ on the test set $\mathcal{D}_{\text{test}}$ can be denoted as:

$$\text{Perf}\,(\mathcal{S}, \mathcal{D}_{\text{test}}) = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{(q,a)\in\mathcal{D}_{\text{test}}} \mathbb{I}(\mathcal{S}(q) = a) \tag{2}$$

Our goal is to select a subset $\mathcal{D}_{\text{selected}} \subseteq \mathcal{D}_{\text{full}}$ from $\mathcal{D}_{\text{full}}$ and make the performance of the student model $\mathcal{S}_{\mathcal{D}_{\text{selected}}}$, distilled using $\mathcal{D}_{\text{selected}}$ on the test set $\mathcal{D}_{\text{test}}$, outperform that of the student model $\mathcal{S}_{\mathcal{D}_{\text{full}}}$ distilled using the full data:

$$\mathcal{D}^*_{\text{selected}} = \arg \max_{\mathcal{D}_{\text{selected}} \subseteq \mathcal{D}_{\text{full}}} \text{Perf}\,(\mathcal{S}_{\mathcal{D}_{\text{selected}}}, \mathcal{D}_{\text{test}}) \tag{3}$$

To achieve the above goal, we designed a four-stage distillation method MoRSD. Its details will be described in the following sections.

### 3.2 Rationale generation

To obtain the dataset for distillation, we adopt the same generation method as in previous studies(Ho et al., 2023). As shown in the upper left of Figure 2, we use a fixed template: "Q: $\langle q_i \rangle$. A: Let's think step by step. $\langle \hat{r}_i \rangle$ Therefore, the answer is $\langle \hat{a}_i \rangle$". By applying this process to all data points in $\mathcal{D}$, we obtain the full dataset $\mathcal{D}_{\text{full}}$ in Eq 1.

### 3.3 Self-evaluation

After building the full dataset $\mathcal{D}_{\text{full}}$ in Section 3.2, we use rationale difficulty (RD) to score each rationale $r_i^j$ in the dataset. RD is a metric based on the perplexity of the student model, where perplexity is the exponential transformation of the normalized Negative Log-Likelihood (NLL), given an input sequence $X = (x_1, x_2, \dots, x_N)$ and a target sequence $Y = (y_1, y_2, \dots, y_M)$, the perplexity can be written as:

$$\text{PPL}(y_j|X) = \exp\left(-\frac{1}{M}\sum_{j=1}^{M}\log\Pr(y_j|x_1,...,x_N,y_{j-1})\right) \tag{4}$$

Since the student model has been pre-trained or supervised-fine-tuned (SFT) using NLL loss on a large corpus of text, its perplexity can indicate the quality of the rationales generated by the teacher. Therefore, we define RD as the ratio of the change in PPL of the student model before and after a given rationale:

$$RD(\hat{r}_i^j, q_i) = \frac{\text{PPL}_{(\theta^S)}(a_i|\hat{r}_i^j, q_i)}{\text{PPL}_{(\theta^S)}(a_i|q_i)}. \tag{5}$$

For rationale $\hat{r}_i^j$, if the student model achieves low $RD(\hat{r}_i^j, q_i)$, it suggests that the rationale is more beneficial for the student in understanding the corresponding question and will be selected in difficulty selection.

### 3.4 Rationale selection

After calculating the RD for each rationale in section 3.3, this section will select a subset $\mathcal{D}_{\text{selected}}$ from the full dataset $\mathcal{D}_{\text{full}}$ based on the accuracy, diversity, and difficulty of the rationale. Therefore, we divide the rationale selection process into three sequential parts: 1) Accuracy Selection, 2) Diversity Selection, and 3) Difficulty Selection.

#### 3.4.1 Accuracy selection

The most important characteristic of rationale is correctness. Different from (Ho et al., 2023; Li et al., 2024), we first divide the rationale into correct and incorrect parts by comparing the final prediction $\hat{a}_i$ of the teacher model with the ground truth $a_i$. We then filter out negative samples to ensure the original dataset meets a given accuracy threshold $\delta$.

Then, we filter the rationales sequentially from the original dataset such that the average accuracy of the filtered dataset $\mathcal{D}_{\text{accurate}}$ reaches $\delta$. The calculation is as follows:

$$\text{Avg Acc} = \frac{1}{|\mathcal{D}_{\text{accurate}}|}\sum_{(\hat{r}_i^j,\hat{a}_i)\in\mathcal{D}_{\text{accurate}}}\text{acc}\left(\hat{r}_i^j,\hat{a}_i\right) \geq \delta$$

#### 3.4.2 Diversity selection

The diversity of rationales is important for distillation performance. However, we found that even with different sampling temperatures, the teacher model often generates similar rationales. To address this, we select diverse rationales by first splitting them into N-grams (N=3 in our experiments). Then, we calculate the pairwise Jaccard similarity between these N-gram sets. For each rationale $r_i^j$, we decompose it into segments $R_i^j$ and use the Jaccard similarity score to compare and identify the most similar rationales.

$$(r_i^m, r_i^n) = \underset{1\leq m,n\leq M, m\neq n}{\arg\max}\frac{|R_i^m \cap R_i^n|}{|R_i^m \cup R_i^n|} \tag{6}$$

We then randomly keep one form the two rationales from Eq. 6 and discard the other. This process repeats until we collect a total of $K$ rationales (set to 6 in our experiments). Afterward, we have a diverse dataset, $\mathcal{D}_{\text{diverse}}$, ready for the final difficulty selection step.

#### 3.4.3 Diffculty selection

After obtaining $\mathcal{D}_{\text{diverse}}$, we need to filter and retain rationales that are helpful for distillation based on RD. As mentioned in section 3.3, rationale with low RD is considered helpful for distillation, so in the difficulty selection, we select the $k$ ($k$ set to 3 in our experiments) samples with the lowest RD in the dataset:

$$\mathcal{D}_{\text{selected}} = \left\{q_i, \{(\hat{r}_i^1,\hat{a}_i^1),(\hat{r}_i^2,\hat{a}_i^2), \quad \dots,(\hat{r}_i^k,\hat{a}_i^k)\}\right\} \tag{7}$$

where $RD\left(\hat{r}_i^1, q_i\right) \leq RD\left(\hat{r}_i^2, q_i\right) \leq \cdots \leq RD\left(\hat{r}_i^k, q_i\right)$, $i = 1, 2, \dots, N^*, j = 1, 2, \dots, M^*$.

| Method | Params | Single Eq | Add Sub | Multi Arith | Strategy QA | GSM8K | SVAMP | Date Understanding | Shuffled Objects |
|---|---|---|---|---|---|---|---|---|---|
| **Random** | - | 0.00 | 0.00 | 0.00 | 50.00 | 0.00 | 0.00 | 17.12 | 33.33 |
| **Teacher: InstructGPT 175B (text-davinci-002)** | | | | | | | | | |
| Zero-shot-CoT | 175B | 82.24 | 78.99 | 78.89 | 53.57 | 40.26 | 64.67 | 73.87 | 50.22 |
| **Student: Flan-T5-{Small, Base, Large, XL}** | | | | | | | | | |
| Vanilla CoT Distillation | 60M | 7.24 | **10.92** | 17.22 | 56.04 | 2.58 | 10.67 | **84.68** | 62.22 |
| | 250M | **9.21** | 10.92 | 21.11 | 60.84 | 4.40 | 12.33 | 84.68 | 67.11 |
| | 780M | 10.52 | 15.13 | 20.00 | 61.72 | 7.12 | 13.67 | 87.39 | 89.33 |
| | 3B | 20.39 | 11.76 | 26.67 | 65.37 | 7.60 | 12.33 | 82.9 | 43.11 |
| Multi-task CoT Distillation | 250M | 5.22 | 8.40 | 8.33 | 52.83 | 6.00 | 2.33 | 80.18 | 31.55 |
| | 780M | 11.89 | 16.81 | 16.81 | 50.09 | 6.36 | 9.00 | 79.23 | 35.96 |
| | 3B | 22.36 | **36.9** | 17.22 | 52.11 | 7.73 | 11.33 | 81.93 | 52.46 |
| MoDE-CoTD | 250M | 5.26 | 7.56 | 13.89 | 56.18 | 6.11 | 5.33 | 85.55 | 35.55 |
| | 780M | 10.52 | 10.92 | 13.89 | 56.47 | 7.28 | 11.33 | **89.19** | 62.22 |
| | 3B | **23.33** | 24.37 | 23.33 | 60.99 | 9.78 | 17.33 | **93.69** | 70.67 |
| MoRSD (ours) | 60M | **9.21** | **10.92** | **22.78** | **60.26** | **6.98** | **11.33** | 82.88 | **83.56** |
| | 250M | **9.21** | **12.61** | **24.44** | **65.65** | **6.98** | **13.67** | **86.49** | **99.56** |
| | 780M | **13.16** | **16.81** | **25.00** | **65.65** | **9.71** | **15.00** | **89.19** | **100.00** |
| | 3B | 21.71 | 24.37 | **31.67** | **65.65** | **10.20** | **23.67** | 91.00 | **100.00** |

Table 1: **MoRSD Performance**. Accuracy (%) of MoRSD and baseline methods on 8 tasks under various settings. **Random** refers to random-guess performance derived based on the number of choices in multi-choice tasks. The best method for each setting is marked in **bold**. For **Zero-shot-CoT**, we use the same prompt setting as (Ho et al., 2023).

### 3.5 Distillation

Then, we use $\mathcal{D}_{selected}$ to fine-tune the student model. Similar to SFT, the objective function of distillation can be written as follows:

$$\mathcal{L}(\theta_S) = -\sum_{r_i \in \mathcal{D}_{\text{selected}}} \mathbf{1}_{(r_i)} \cdot \log \Pr\left(a_i, \hat{r_i} \mid q_i; \theta_S\right) \quad (8)$$

The final distilled student model $\mathcal{D}_{\text{selected}}$ is used to verify the final performance according to Eq 2.

## 4 Experiment

### 4.1 Task and Datasets

Experiments were conducted on seven datasets related to three tasks: mathematical reasoning, question answering, and temporal/spatial reasoning. Including StrategyQA (Geva et al., 2021) for commonsense reasoning, Addsub (Hosseini et al., 2014), Multiarith (Roy and Roth, 2015), SVAMP (Patel et al., 2021), SingleEq (Koncel-Kedziorski et al., 2015) and GSM8K (Cobbe et al., 2021) for arithmetic math inference and Date Understanding (Srivastava et al., 2023), Tracking Shuffled Objects (Srivastava et al., 2023) for temporal/spatial reasoning. The details on partition training, testing sets, and other specificities are provided in the Appendix A.

### 4.2 Baseline

We provide a comparison of MoRSD (ours) with three baseline methods:

• **Vanilla CoT Distillation** (Ho et al., 2023), where the student model is directly fine-tuned on the teacher-generated CoT rationales without additional selection or filtering.

• **Multi-task CoT Distillation** (Li et al., 2024), where the student model is fine-tuned on a combined dataset from multiple reasoning tasks.

• **MoDE-CoTD** (Li et al., 2024), where the rationales from different tasks are distilled into separate LoRA modules, enabling cross-task collaboration through task-specific parameter adaptation.

• **MCC-KD** (Chen et al., 2023), which improves reasoning consistency by generating multiple rationales per question and minimizing bidirectional KL-divergence between their answer distributions.

• **Mentor-KD** (Lee et al., 2024), which uses a task-specific mentor model to enrich the distillation set with CoT annotations and soft labels, addressing data quality and label scarcity.

Table 2: **Performance of MoRSD and baselines across two student models on four tasks.** Best results for each student model are in **bold**.

| Method | Student | Strategy QA | SVAMP | Date Understanding | Shuffled Object | Average |
|---|---|---|---|---|---|---|
| MCC-KD | **FlanT5-Small** | 58.37 | 10.00 | 81.98 | 43.11 | 48.37 |
| MENTOR-KD | **FlanT5-Small** | 59.97 | 10.67 | 83.78 | 82.67 | 59.27 |
| MORSD (OURS) | **FlanT5-Small** | **61.35** | **12.33** | **84.43** | **84.69** | **60.70** |
| MCC-KD | **FlanT5-Base** | 64.92 | 12.00 | 85.59 | 69.78 | 58.07 |
| MENTOR-KD | **FlanT5-Base** | 65.21 | 11.33 | **87.39** | 93.78 | 64.43 |
| MORSD (OURS) | **FlanT5-Base** | **65.72** | **14.28** | 87.04 | **99.62** | **66.67** |

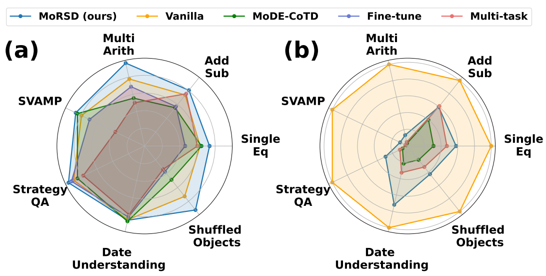


Figure 3: **Comparison of the performance and the rationale usage**.

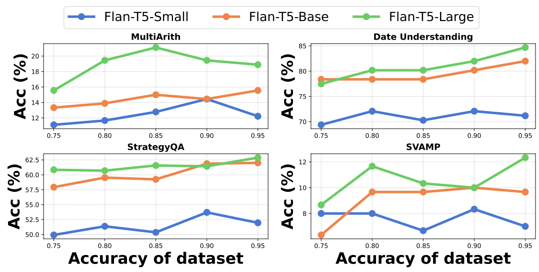


Figure 4: **Effect of dataset accuracy**. The performances of MoRSD on the MultiArith, Date Understanding, StrategyQA and SVAMP datasets with different correctness rates of the teacher generated rationales.

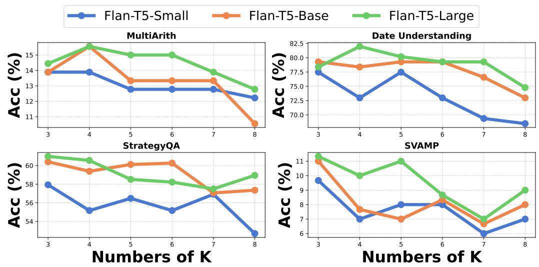


Figure 5: **Effect of rationale diversity**. The performance of MoRSD on four test sets with different rationale diversities.

### 4.3 Teacher and Student Models

For the teacher models, we use GPT-3 175B (Brown et al., 2020), accessed via the OpenAI API, with *text-davinci-002* (Ouyang et al., 2022) as the default model unless otherwise specified. We employ the instruction-tuned versions of T5 for the student models, specifically Flan-T5-{Small, Base, Large} (Chung et al., 2022).

## 5 Results

In this section, we report the performance of our MoRSD and baseline methods on 7 benchmarks. We compare our approach with baselines of different model sizes. The performance on the test set demonstrates the effectiveness of our approach, showing that our method achieves better performance with fewer samples.

### 5.1 MoRSD outperforms baselines across different student models

The results in Table 1 and Table 2 show that **MoRSD consistently outperforms strong baselines** across various student model sizes and reasoning tasks. On **Flan-T5-Small**, MoRSD notably improves results on challenging datasets such as **SVAMP** and **Tracking Shuffled Objects**, achieving **11.33%** on SVAMP (+3.73% over MoDE-CoTD) and **83.56%** on Tracking Shuffled Objects, surpassing MoDE-CoTD (62.22%) and Multi-task CoT (31.55%). These improvements are obtained with **fewer rationales**, highlighting the effectiveness of selective rationale filtering over data quantity.

Compared to **multi-task and consistency-based methods** like **MCC-KD** and **Mentor-KD**, MoRSD achieves **comparable or better performance**. On Flan-T5-Small, it reaches an **average accuracy of 59.51%**, slightly above Mentor-

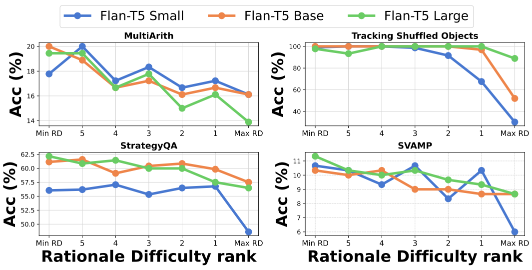


Figure 6: **Effect of rationale difficulty**.The performance of MoRSD using different samples selected by RD among four test sets

KD (59.27%) and notably higher than MCC-KD (48.37%), demonstrating that **effective rationale selection can boost performance without extra supervision**.

As the **student model scales up**, MoRSD **continues to outperform baselines**. On **Flan-T5-Base**, it achieves the **highest average accuracy of 66.34%**, exceeding Mentor-KD (64.43%) and MCC-KD (58.07%). Notably, MoRSD achieves **near-perfect accuracy** on temporal and spatial reasoning tasks such as **Tracking Shuffled Objects (99.56%)** and **Date Understanding (86.49%)**, indicating strong generalization.

## 5.2 Effect of rationale correctness and diversity

To assess how rationale accuracy affects distillation, we varied dataset accuracy and measured student performance. As shown in Figure 4, distillation improves with higher accuracy, but gains plateau beyond a certain threshold. This indicates that accuracy is crucial at lower levels, while its marginal benefit diminishes as it increases.

The diversity of the rationale is also vital for distillation. To measure the degree of diversity among rationales, we use the number of rationales remaining after the Jaccard similarity filtering to measure the diversity of the dataset. In simple terms, a smaller number of remaining rationales after filtering indicates a higher level of diversity in the dataset. As illustrated in Figure 5, the performance of MoRSD exhibits a corresponding improvement with increasing diversity among the rationales, as observed in all four different test sets.

## 5.3 Effect of rationale difficulty

To verify the effect of the rationale difficulty (RD) on distillation performance, we conducted experiments using samples of varying sizes selected after sorting based on RD. As illustrated in Figure 6, the distillation performance of the student model improves as the RD of the selected data decreases, achieving optimal performance when the RD is at its smallest. This trend is consistent across multiple test sets, including StrategyQA and Tracking Shuffled Objects, demonstrating that lower RD values correlate with more effective distillation outcomes. The results underscore the efficacy of the proposed RD indicator in identifying and prioritizing data that is most beneficial for the distillation process. This finding highlights the importance of RD in enhancing the overall performance of the student model by focusing on the most informative and manageable rationales.

## 5.4 Ablation study

In this section, we conduct an ablation study on the Flan-T5-Small model to assess the contributions of accuracy, diversity, and difficulty selection in MoRSD. As shown in Table 3, removing any component leads to notable performance drops. Accuracy selection is critical—its removal causes large degradations on tasks like **SingleEq** ($-35.7\%$) and **SVAMP** ($-38.1\%$). Diversity selection is especially important for reasoning-heavy tasks such as **MultiArith** ($-31.2\%$) and **Tracking Shuffled Objects** ($-44.3\%$), helping reduce redundancy. Difficulty selection prioritizes informative rationales, and its absence also leads to significant drops, including $-44.0\%$ on **SingleEq** and $-26.9\%$ on **SVAMP**. These results indicate that each selection stage plays a distinct and complementary role in improving distillation effectiveness. Overall, all three components are essential for maximizing student performance.

## 5.5 Analyse of selected rationale

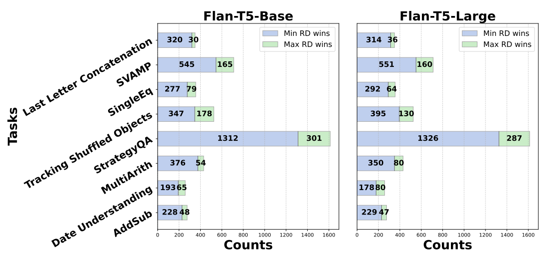


Figure 7: **Effect of selected rationale**. The ChatGPT API was used as a referee, prompted to compare two rationales and rate them on a scale of 1 to 10. Each rationale pair (maximum RD and minimum RD) was judged twice to avoid position bias, with the rationale positions swapped in each evaluation.

| Method | Single Eq | Add Sub | Multi Arith | Strategy QA | SVAMP | Date Understanding | Shuffled Objects |
|---|---|---|---|---|---|---|---|
| MoRSD | 9.21 | 10.92 | 22.78 | 60.26 | 11.33 | 82.88 | 82.22 |
| w/o Accuracy Sel. | 5.92 -3.29 | 10.08 -0.84 | 15.00 -7.78 | 57.21 -3.05 | 5.67 -5.66 | 74.77 -8.11 | 89.33 +7.11 |
| w/o Diversity Sel. | 9.21 -0.00 | 10.92 -0.00 | 15.00 -7.78 | 59.64 -0.62 | 4.67 -6.66 | 82.88 -0.00 | 67.56 -14.66 |
| w/o Difficulty Sel. | 1.97 -7.24 | 8.40 -2.52 | 15.56 -7.22 | 60.26 -0.00 | 7.33 -4.00 | 76.58 -6.30 | 82.22 -0.00 |

Table 3: **Ablation study on Flan-T5-Small**. Results of ablation study about Accuracy selection, Diversity selection, and Difficulty selection on test sets.

In order to compare the quality of rationales screened by different methods, we introduced the ChatGPT API as a referee to further explore the characteristics of different rationales selected using RD. By stitching different rationales together and prompting the referee to judge which of the two is better and give them a score of 1-10, we visualized these results as the winning frequency of those selected with the minimum RD and the maximum RD. As presented in Figure 7, to avoid possible bias of the judges due to the position of the rationale in the prompt, we judged each maximum RD-minimum RD pair twice and exchanged the position of the rationale in the prompt in each judgment. From the results, we can conclude that the quality of rationales with lower RD attributes is higher than those with higher RD attributes on all datasets. This further proves the effectiveness of the RD in selecting high-quality rationales.

### 5.6 Effect of negative rationale

To further examine the role of imperfect rationales and supervision signals, we conduct an ablation study across three labeling strategies, as summarized in Table 4. The Positive-only setting relies exclusively on gold rationales and answers, achieving reasonable performance but with limited diversity. Our default MoRSD configuration, which selects rationales based on Rationale Difficulty (RD) while always supervising with the gold answer, yields consistent improvements across all tasks (e.g., +0.8 on StrategyQA and +1.3 on Date Understanding) and achieves the best overall average (66.4). This confirms that RD-based selection enhances both quality and diversity of rationales without sacrificing correctness. In contrast, replacing gold answers with teacher-predicted labels substantially degrades performance (e.g., a 3.9 drop on SVAMP and 4.0 on Date Understanding), highlighting the necessity of grounding training supervision in correct labels. These results validate our design choice: selectively retaining imperfect rationales improves robustness and data efficiency, but correctness of the final answer supervision remains critical.

## 6 Discussion

### 6.1 Inclusion of Negative Rationales

An important design decision in MoRSD concerns the treatment of rationales that do not lead to correct teacher predictions. While conventional approaches often discard such negative rationales entirely, we deliberately retain a subset that passes our Rationale Difficulty (RD) filter. Specifically, a rationale is preserved if it reduces the student's perplexity in predicting the ground-truth answer, even when the intermediate reasoning is partially incorrect. This choice is motivated by prior findings that structural patterns in CoT traces can facilitate learning even when their semantic content is imperfect.

### 6.2 Generalization Across Tasks and Domains

While MoRSD demonstrates consistent improvements on diverse reasoning benchmarks, its current evaluation scope remains primarily within math and structured reasoning tasks. An open question is how well the rationale selection paradigm generalizes to other domains, where the structure of rationales may differ significantly from mathematical derivations. Moreover, in multilingual or cross-domain scenarios, the reliability of perplexity-based Rationale Difficulty (RD) as a selection signal could be weakened, since student models may not share the same linguistic or distributional priors as their teachers. Future work could explore extending MoRSD to multilingual reasoning tasks, domain-adaptation settings, thereby testing whether the principle of less but better rationales remains universally effective beyond the current experimental scope.

| **Method Setting** | **Label Type** | SVAMP | Strategy QA | Date Understanding | Shuffled Objects |
|---|---|---|---|---|---|
| Positive-only | Ground-truth | 13.2 | 64.9 | 85.2 | 98.4 |
| MoRSD | Ground-truth | 13.7 | 65.7 | 86.5 | 99.6 |
| MoRSD + Predict label | Teacher prediction | 10.2 | 61.7 | 82.0 | 96.1 |

Table 4: Ablation study on labeling strategies in MoRSD (Flan-T5-Base). **Positive-only** uses only gold rationales and answers, **MoRSD** applies RD-based rationale selection with gold supervision and achieves the best overall accuracy, while **MoRSD + Predict label** replaces gold labels with teacher predictions and suffers clear degradation. Results show that RD-based selection improves data efficiency, but gold answer supervision is crucial.

### 6.3 Considerations for RD-Based Selection

In MoRSD, Rationale Difficulty (RD) serves as the central criterion for rationale selection by measuring the student's perplexity reduction on gold answers. While effective, RD captures only part of rationale quality: it reflects token-level uncertainty but may not align with logical soundness or pedagogical value. In addition, distributional mismatch across domains or languages could further reduce its reliability. Future work could extend MoRSD by integrating RD with complementary signals—such as process reward models or structural coherence metrics—to achieve more robust rationale evaluation.

## 7 Conclusion

In this work, we propose **MoRSD**, an efficient CoT distillation method that enhances the performance of small language models using fewer rationales. By introducing a self-guided Rationale Difficulty metric, MoRSD enables the autonomous selection of high-quality rationales, effectively addressing challenges related to the rationale quality. Experiments across seven datasets demonstrate an average accuracy improvement of 4.6% over the baseline. MoRSD outperforms full dataset distillation with a small, tailored set of rationales, providing a robust solution for efficient CoT distillation and advancing knowledge transfer in a more efficient manner.

## Limitations

Although MoRSD achieves significant improvements on the Flan-T5 series but is not universally applicable. First, the selection based on rationale difficulty requires the student model to have a basic capability, making it unsuitable for models without fine-tuning. Applying MoRSD to such models would require instruction fine-tuning, increasing computational costs. Second, selecting high-quality rationales requires filtering a large dataset from the teacher model, matching the computational cost of traditional CoT distillation. Future work could focus on efficient rationale generation. Moreover, the selection method relies on the student model's perplexity, which may introduce bias due to its parameter size. While small RD identifies most high-quality samples, it cannot exclude all low-quality rationales, potentially affecting distillation results.


## Acknowledgments

This work is supported by the National Science and Technology Major Program (2024ZD01NL00101), the Major Key Project of PCL (PCL2025A03), the Natural Science Foundation of China (62506182, 62276082), National Key RD Program of China (2023YFC3502900), Shenzhen Science and Technology Research and Development Fund (KJZD20240903102802003), Shenzhen Soft Science Research Program Project (RKX20220705152815035), Shenzhen Science and Technology Research and Development Fund for Sustainable Development Project (GXWD20231128103819001, KCXFZ20201221173613036, 20230706140548006) and Guangdong Provincial Key Laboratory (2023B1212060076).


## References


Mohamad Ballout, Ulf Krumnack, Gunther Heidemann, and Kai-Uwe Kühnberger. 2024. Efficient knowledge distillation: Empowering small language models with teacher model insights. In *Natural Language Processing and Information Systems: 29th International Conference on Applications of Natural Language to Information Systems, NLDB 2024, Turin, Italy, June 25–27, 2024, Proceedings, Part I*, page 32–46, Berlin, Heidelberg. Springer-Verlag.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Dan Busbridge, Amitis Shidani, Floris Weers, Jason Ramapuram, Etai Littwin, and Russ Webb. 2025. Distillation scaling laws. *Preprint*, arXiv:2502.08606.

Hongzhan Chen, Siyue Wu, Xiaojun Quan, Rui Wang, Ming Yan, and Ji Zhang. 2023. Mcc-kd: Multi-cot consistent knowledge distillation. *Preprint*, arXiv:2310.14747.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. 2024. Alpagasus: Training a better alpaca model with fewer data. In *The Twelfth International Conference on Learning Representations*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, and et al. Gaurav Mishra. 2022. Palm: Scaling language modeling with pathways. *Preprint*, arXiv:2204.02311.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

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

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

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, and Z. F. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, and Bingxuan Wang. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023. Specializing smaller language models towards multi-step reasoning. *Preprint*, arXiv:2301.12726.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Preprint*, arXiv:2101.02235.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. Large language models are reasoning teachers. *Preprint*, arXiv:2212.10071.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack William Rae, and Laurent Sifre. 2022. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*.

Mohammad Javad Hosseini, Hannaneh Hajishirzi, Oren Etzioni, and Nate Kushman. 2014. Learning to solve arithmetic word problems with verb categorization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 523–533, Doha, Qatar. Association for Computational Linguistics.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander 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. *Preprint*, arXiv:2305.02301.

Bumjun Kim, Kunha Lee, Juyeon Kim, and Sangam Lee. 2024. Small language models are equation reasoners. *Preprint*, arXiv:2409.12393.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. *Preprint*, arXiv:2205.11916.

Rik Koncel-Kedziorski, Hannaneh Hajishirzi, Ashish Sabharwal, Oren Etzioni, and Siena Dumas Ang. 2015. Parsing algebraic word problems into equations. *Transactions of the Association for Computational Linguistics*, 3:585–597.

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

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, Toronto, Canada. Association for Computational Linguistics.

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, Torino, Italia. ELRA and ICCL.

Dheeraj Mekala, Alex Nguyen, and Jingbo Shang. 2024. Smaller language models are capable of selecting instruction-tuning training data for larger language models. *Preprint*, arXiv:2402.10430.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *Preprint*, arXiv:2501.19393.

OpenAI. 2024. Learning to reason with llms.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, and Diogo Almeida. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online. Association for Computational Linguistics.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. *Preprint*, arXiv:1906.02361.

Subhro Roy and Dan Roth. 2015. Solving general arithmetic word problems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1743–1752, Lisbon, Portugal. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, and Adam R. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Preprint*, arXiv:2206.04615.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, Weixin Liu, Zhihua Wu, Weibao Gong, Jianzhong Liang, Zhizhou Shang, Peng Sun, Wei Liu, Xuan Ouyang, Dianhai Yu, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *Preprint*, arXiv:2107.02137.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Peifeng Wang, Aaron Chan, Filip Ilievski, Muhao Chen, and Xiang Ren. 2023a. Pinto: Faithful language reasoning using prompt-generated rationales. *Preprint*, arXiv:2211.01562.

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. 2023b. Scott: Self-consistent chain-of-thought distillation. *Preprint*, arXiv:2305.01879.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *Preprint*, arXiv:2409.12122.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025. Limo: Less is more for reasoning. *Preprint*, arXiv:2502.03387.

Jiahao Ying, Mingbao Lin, Yixin Cao, Wei Tang, Bo Wang, Qianru Sun, Xuanjing Huang, and Shuicheng Yan. 2024. Llms-as-instructors: Learning from errors toward automating model improvement. *Preprint*, arXiv:2407.00497.

Yuanhao Yue, Chengyu Wang, Jun Huang, and Peng Wang. 2024. Distilling instruction-following abilities of large language models with task-aware curriculum planning. *Preprint*, arXiv:2405.13448.

Ruichen Zhang, Hongyang Du, Yinqiu Liu, Dusit Niyato, Jiawen Kang, Sumei Sun, Xuemin Shen, and H. Vincent Poor. 2024a. Interactive ai with retrieval-augmented generation for next generation networking. *IEEE Network*, 38(6):414–424.

Yifei Zhang, Bo Pan, Chen Ling, Yuntong Hu, and Liang Zhao. 2024b. Elad: Explanation-guided large language models active distillation. In *Findings of the Association for Computational Linguistics ACL 2024*, page 4463–4475. Association for Computational Linguistics.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment. *Preprint*, arXiv:2305.11206.

# A Appendix

## A.1 Datasets

A summary of the datasets used in our experiments, along with their original licenses, can be found in Appendix Table 5. We utilize the 7 datasets from (Kojima et al., 2023) to evaluate reasoning performance.

| Dataset | Training Samples | Test Samples | Data Split | License |
|---|---|---|---|---|
| SingleEq | 356 | 152 | 70:30 | None |
| AddSub | 276 | 119 | 70:30 | Unspecified |
| MultiArith | 420 | 180 | 70:30 | Unspecified |
| SVAMP | 700 | 300 | 70:30 | MIT |
| Date Understanding | 258 | 111 | 70:30 | Apache-2.0 |
| Tracking Shuffled Objects | 525 | 225 | 70:30 | Apache-2.0 |
| StrategyQA | 1603 | 687 | 70:30 | Apache2.0 |

Table 5: Description of datasets used in our study.

## A.2 Experimental details

All experiments were conducted on a cluster of NVIDIA V100 GPUs. We strictly controlled the hyperparameters for all datasets. For each experiment, we used a batch size of 8 and a maximum of 10,000 steps, which was found to be sufficient for the test accuracy to plateau. We report the best accuracy achieved within these 10,000 steps.

## A.3 KDE visualization of API scores

In Section 5.5, we used the ChatGPT-API to score rationales on a scale of 1 to 10 and employed KDE to visualize the score distributions for rationales selected by different methods. The KDE distributions for rationales selected via the minimum RD approach (red curves) show distinct advantages across tasks, with scores concentrated between 6 and 8, indicating higher and more consistent quality compared to other methods. The mean values of these distributions (dashed red lines) are consistently higher than those of maximum RD rationales (dashed blue lines), further supporting the superiority of the minimum RD method.

However, tasks like StrategyQA and Tracking Shuffled Objects exhibit longer tails in the minimum RD distributions, indicating a small proportion of lower-quality outliers. Despite this variability, the minimum RD method generally selects higher-quality rationales, making it a more effective approach for ensuring better overall quality in most cases.

## A.4 RD and length

Figure 9 illustrates the relationship between rationale length and tokenized rationale length for different model sizes of Flan-T5 {small, base, large}. As the rationale length increases, the tokenized rationale length grows correspondingly, with a more pronounced increase observed in larger model versions. For the Flan-T5-small model, the rate of growth is moderate, indicating that smaller models require fewer tokens for shorter rationales. In contrast, the Flan-T5-base model shows a steeper increase in tokenized length as rationale length grows, reflecting its enhanced capacity to handle more complex reasoning. The Flan-T5-large model exhibits the most significant acceleration in tokenized rationale length, suggesting that larger models, with their greater capacity, demand significantly more tokens for longer rationales. This trend highlights the models' scaling behaviour, where larger models can handle more extensive rationales, necessitating an increase in the number of tokens for effective representation. Overall, the results underscore the positive correlation between rationale length and tokenized length across all model sizes, with the rate of increase being more pronounced in larger models.

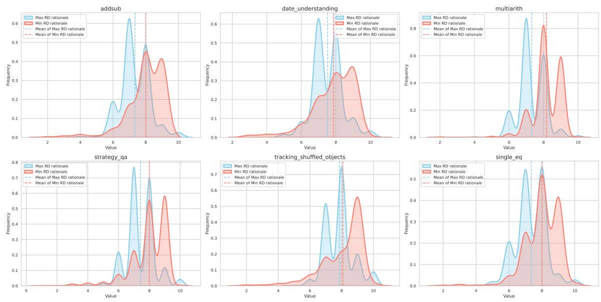

Figure 8: **KDE plot of scored selected rationale.** Kernel Density Estimation (KDE) plot, where the ChatGPT API is employed as a referee to investigate the characteristics of various rationales selected through RD. By combining different reasons and assigning them scores ranging from 1 to 10..

## A.5 Transferability of rationale selected by RD

To verify whether the RD calculated by different models can also improve the distillation performance on other models, we use Flan-T5-Small, Base, Large and the larger LLamA2-7b-hf to calculate their respective RDs and use them to fine-tune the smaller Flan-T5-Small and use the RD calculated by Flan-T5-Small to fine-tune the larger Flan-T5-Base model. The RD transferability analysis and wilcoxon signed-rank test in Table 7 reveals that RD transfer from different models (Flan-T5

| Prompt for Performance Evaluation | |
|---|---|
| **System Prompt** | You are a helpful and precise assistant for checking the quality of the rationale based on a given question. |
| **Task Discribe** | We would like to request your feedback on the performance of two rationales in response to the question displayed above. Please rate the rationales. Each rationale receives an overall score on a scale of 1 to 10, where a higher score indicates better overall performance. Please first output a single line containing only two values indicating the scores for rationale 1 and rationale 2, respectively. The two scores are separated by a space. In the subsequent line, please provide a comprehensive explanation of your evaluation and fully compare the quality of the two rationales, avoiding any potential bias and ensuring that the order in which the rationale was presented does not affect your judgment. |
| **Prompt** | `[Question] {question}  [The Start of Rationale1] {rationale_1} [The End of Rationale1]  [The Start of Rationale2] {rationale_2} [The End of Rationale2]  [System] {TASK_DISCRIBE}` |

Table 6: The prompt we used to request ChatGPT to evaluate the rationales.

| Train model / Cal RD model | AddSub | | | | SingleEq | | | | StrategyQA | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Flan-T5 Small** | **p-value** | **Flan-T5 Base** | **p-value** | **Flan-T5 Small** | **p-value** | **Flan-T5 Base** | **p-value** | **Flan-T5 Small** | **p-value** | **Flan-T5 Base** | **p-value** |
| Flan-T5 Small | 4.77 | – | -1.13 | $3.8e^{-7}$ | 3.28 | – | -0.72 | 0.14 | 52.76 | – | -1.39 | 0.02 |
| Flan-T5 Base | +1.04 | $1.5e^{-5}$ | 5.64 | – | +0.22 | 0.54 | 4.72 | – | +1.33 | $6.2e^{-4}$ | 58.39 | – |
| Flan-T5 Large | +1.36 | $2.3e^{-6}$ | +1.67 | $6.4e^{-9}$ | +0.16 | 0.45 | +0.19 | 0.97 | +0.58 | 0.02 | +0.55 | $6.1e^{-7}$ |
| LLamA2-7B-hf | +1.40 | 0.001 | +1.28 | 0.141 | +0.66 | 0.062 | +0.87 | 0.012 | +0.39 | 0.223 | +0.80 | 0.082 |

Table 7: **Transferability analyse for RD**. Flan-T5-{Small, Base, Large} and LLaMA2-7B are used to calculate their RDs, which are then used to distill Flan-T5-Small. Conversely, the RD from Flan-T5-Small, Large and LLaMA2-7B is used to distill Flan-T5-Base.

variants and LLaMA2-7B) improves performance more on simpler tasks than on complex ones. For tasks like AddSub and SingleEq, RD transfer from Flan-T5 Base and Large results in notable improvements, with Flan-T5 Large showing increases of 1.36% in AddSub (p-value = 0.009) and 1.68% in SingleEq (p-value = 0.001). However, the gains are minimal for the more complex StrategyQA task, with Flan-T5 Large only improving performance by 0.58% (p-value = 0.269). Overall, the transfer of reasoning capabilities through RD (Rationale Distillation) proves to be more effective for relatively simple tasks, where smaller models benefit significantly from the distillation process. In contrast, the impact of using larger models in such tasks tends to be less pronounced.

### A.6 Prompt for evaluation

In this section, we provide the detailed prompt we used for evaluating the performance of two rationales for the same instruction as shown in Table 6

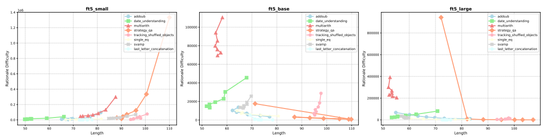

Figure 9: Comparison of RD Lengths

### A.7 Pattern Characteristics Comparison of Selected rationale

In order to better compare the quality difference between the maximum RD and minimum RD rationales, we use ChatGPT's API to compare them and give an explanation. As shown in Table 8 9 and 10, the primary advantage of the rationale with min RD over the rationale with max RD is its more detailed and coherent reasoning process. It clearly breaks down each step of the reasoning, providing explicit explanations for how the final conclusion is reached, which enhances both transparency and logical rigor. By systematically deconstructing the problem, the rationale with min RD allows readers to more easily follow the reasoning flow. In contrast, the rationale with max RD, while more concise, may lack sufficient detail and explanation, potentially causing confusion. As a result, the rationale with min RD generally leads to a clearer understanding of the reasoning process.

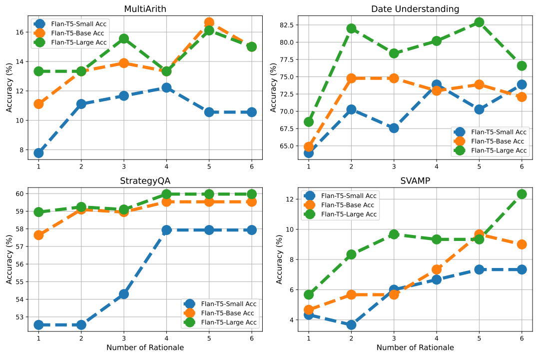


Figure 10: **Effect of the number of rationales selected by RD**. The performance of MoRSD with different numbers of rationales is evaluated on the Date Understanding, StrategyQA, SVAMP, and MultiArith datasets.

### A.8 Effect of the number of rationales selected by RD

We also examine the performance of MoRSD with varying numbers of rationales on Date Understanding, StrategyQA, Tracking Shuffled Objects, and MultiArith, respectively. The student model employed in these experiments is Flan-T5-Small. As depicted in Figure 10, we observe that as the number of rationales increases, the model's performance on all four datasets improves correspondingly. Specifically, when the number of rationales is increased from 1 to 4, performance is significantly enhanced on both datasets. However, the performance gains become less pronounced when the number is further increased from 4 to 6. Therefore, considering computational efficiency, we use 6 rationales in our experiments.

| | |
|---|---|
| **AddSub** | |
| **Question** | There were 28 bales of hay in the barn . Tim stacked bales in the barn today . There are now 54 bales of hay in the barn . How many bales did he store in the barn ? |
| **Maximum RD** | There were 28 bales of hay in the barn. Tim stacked bales in the barn today. There are now 54 bales of hay in the barn. How many bales did he store in the barn? He must have stored 26 bales. –> 26 |
| **Minimum RD** | We know that there were 28 bales of hay in the barn. We know that Tim stacked bales in the barn today. We know that there are now 54 bales of hay in the barn. So, how many bales did Tim store in the barn? We can solve this problem by using basic algebra. First, we need to create a variable to represent the number of bales that Tim stored in the barn. We will use the letter "$x$" to represent this variable. Next, we need to create an equation that represents the information –> 26 |
| **API judgement** | 7 9 Rationale 1: Score: 7 Strengths: - Provides a direct and concise answer to the question by calculating the difference between the initial and final number of bales. Weaknesses: - Lacks explanation or reasoning behind the calculation. - Could benefit from more context or steps to show the process of arriving at the solution. Rationale 2: Score: 9 Strengths: - Demonstrates a clear step-by-step approach using algebra to solve the problem. - Provides thorough explanations and sets up an equation to represent the situation accurately. - Offers a more detailed and structured response compared to Rationale 1. Weaknesses: - Could further enhance clarity by breaking down complex steps into simpler terms for better understanding. Comparison: Rationale 2 outperforms Rationale 1 due to its comprehensive and structured approach. In conclusion, Rationale 2 offers a higher quality response by incorporating algebraic methods and providing detailed reasoning, earning it a higher score in this evaluation. |
| **Date Understanding** | |
| **Question** | The deadline is Jun 1, 2021, which is 2 days away from now. What is the date 10 days ago in MM/DD/YYYY? Which choice is true? Answer choices: (A) 06/24/2021, (B) 04/06/2021, (C) 04/29/2021, (D) 05/21/2021, (E) 05/20/2021. |
| **Maximum RD** | The deadline is Jun 1, 2021, which is 2 days away from now. So, 10 days ago would be May 21, 2021. Therefore, the answer is (D). –> E |
| **Minimum RD** | Jun 1, 2021 is 2 days away from now. 2 days ago would be May 30, 2021. 10 days ago from May 30, 2021 would be May 20, 2021. Therefore, the answer is (E) 05/20/2021. –> E |
| **API judgement** | 7 9 Rationale 1: Score: 7 Explanation: This rationale correctly identified that the deadline is on Jun 1, 2021, which is 2 days away from the current date. The final answer choice selected was incorrect, but the reasoning process was logical and clear. Rationale 2: Score: 9 Explanation: This rationale not only correctly identified the current date and calculated 10 days ago accurately but also provided a step-by-step breakdown of the calculations. Comparison: Rationale 2 outperformed Rationale 1 due to its more detailed and coherent explanation. Rationale 2 not only provided the correct answer but also presented the steps involved in reaching that answer clearly. Rationale 1, although accurate in its calculation, lacked the detailed breakdown that Rationale 2 offered. |

Table 8: Example of Selected rationale and API judgement.

| Multiarith | |
|---|---|
| **Question** | For Halloween Debby and her sister combined the candy they received. Debby had 32 pieces of candy while her sister had 42. If they ate 35 pieces the first night, how many pieces do they have left? |
| **Maximum RD** | Debby took 24 pictures at the zoo, and then she took 12 more at the museum, so she took a total of 36 pictures. If she later deleted 14 pictures, she would have 36 - 14, or 22 pictures. –> 22 |
| **Minimum RD** | Debby took 24 pictures at the zoo and 12 at the museum. That's a total of 36 pictures. If she later deleted 14 of the pictures, she would still have 22 pictures. –> 22 |
| **API judgement** | 7 9 Rationale 1: The rationale correctly identifies the total number of pictures Debby took (36) and subtracts the number she deleted (14) to find the remaining amount. The explanation is clear and concise, making it easy to follow the logic. However, it could be improved by explicitly stating the final answer instead of just showing the calculation. Rationale 2: This rationale also correctly calculates the total number of pictures, mentions the deletion process, and provides the final answer. In addition to the clear logic, it includes a better structure with separate lines for the calculations, making it easier to read and understand. Overall, Rationale 2 is rated higher due to its better organization and presentation of information. While Rationale 1 is also accurate, it lacks the clarity and structure of Rationale 2. |
| **Strategy QA** | |
| **Question** | Would toast for a vegan have margarine instead of butter? |
| **Maximum RD** | Toast is bread that has been heated until it is crisp. Bread is made from flour, water, yeast, and salt. Flour is made from wheat. So, no, toast would not have margarine instead of butter. –> Yes |
| **Minimum RD** | First, what is veganism? Veganism is the practice of abstaining from the use of animal products, particularly in diet, and an associated philosophy that rejects the commodity status of animals. So, a vegan does not eat any animal products. This includes, but is not limited to, meat, poultry, fish, dairy, eggs, and honey. Now, let's think about toast. Toast is a quick and easy breakfast food that can be made with just a few ingredients. Bread and butter are the two main ingredients in toast. Bread is vegan. There are many –> Yes |
| **API judgement** | 6 9 Rationale 1 provides a basic explanation of veganism and the restrictions it entails in terms of avoiding animal products. It correctly points out that a vegan would not consume butter due to it being an animal product. However, the rationale lacks a direct answer to the question regarding margarine as an alternative for vegans. On the other hand, Rationale 2 concisely explains the ingredients of toast and clearly states that toast would not have margarine instead of butter. It directly addresses the question by explaining that toast consists of basic plant-based ingredients and does not require butter or margarine. Therefore, Rationale 2 scores higher for providing a more straightforward and accurate response to the given question compared to Rationale 1. |

Table 9: Example of Selected rationale and API judgement.

| | |
|---|---|
| **Single Eq** | |
| **Question** | Sally had 760 quarters in her bank. She spent 418 of her quarters. How many quarters does she have now ? |
| **Maximum RD** | Sally has 760 quarters in her bank. She spends 418 of her quarters. That means she has lost 418 quarters. So we have to subtract 418 from 760. 760-418=342 Therefore, Sally has 342 quarters left in her bank. –> 342.0 |
| **Minimum RD** | Sally had 760 quarters in her bank. She spent 418 of her quarters. This means that Sally now has 760 - 418 = 342 quarters in her bank. –> 342.0 |
| **API judgement** | 5 7 Rationale 1: The rationale correctly identifies the initial number of quarters Sally had, the amount she spent, and calculates the remaining quarters accurately. Rationale 2: This rationale also correctly identifies the initial and remaining quarters but presents the solution more concisely compared to Rationale 1. Overall, Rationale 2 is rated higher (7) due to its clarity and conciseness, while Rationale 1 is rated at 5 as it provides the correct solution but in a slightly more convoluted manner. Both solutions correctly arrive at the answer, but Rationale 2 presents the information in a more straightforward and efficient way. |
| **Tracking shuffled objects** | |
| **Question** | Alice, Bob, and Claire are playing a game. At the start of the game, they are each holding a ball: Alice has a black ball, Bob has a brown ball, and Claire has a blue ball. As the game progresses, pairs of players trade balls. First, Bob and Claire swap balls. Then, Alice and Bob swap balls. Finally, Claire and Bob swap balls. At the end of the game, Bob has the Which choice is true? Answer choices: (A) black ball, (B) brown ball, (C) blue ball. |
| **Maximum RD** | First, Bob and Claire swap balls. This means that Bob now has the blue ball, and Claire has the brown ball. Then, Alice and Bob swap balls. This means that Alice now has the brown ball, and Bob has the black ball. Finally, Claire and Bob swap balls. This means that Claire now has the black ball, and Bob has the blue ball. At the end of the game, Bob has the blue ball. Therefore, the correct answer is (C). –> B |
| **Minimum RD** | output:" After Bob and Claire swap, Claire has the brown ball and Bob has the blue ball. Then, after Alice and Bob swap, Alice has the brown ball and Bob has the black ball. Finally, after Claire and Bob swap, Claire has the black ball and Bob has the brown ball. Therefore, answer choice B, Bob has the brown ball, is correct. –> B |
| **API judgement** | 7 9 Rationale 1 provides a clear and logical explanation of how the ball swapping between the players leads to Bob ending up with the blue ball. The rationale demonstrates a good understanding of the question and effectively leads to the correct answer. Rationale 2 also correctly follows the ball swapping sequence and provides a coherent explanation, leading to the conclusion that Bob ends up with the brown ball. The explanation is concise and clear, with proper transitions between the steps. This rationale also exhibits a good grasp of the question and leads to the correct answer smoothly. When comparing the two rationales, Rationale 2 receives a higher score due to its more concise and straightforward explanation. Additionally, Rationale 2 has a lower perplexity value, indicating more concise reasoning. However, both explanations ultimately reach the correct answer and demonstrate a solid understanding of the game’s mechanics. |

Table 10: Example of Selected rationale and API judgement.