

# 000 WHERE DID THIS SENTENCE COME FROM? TRACING 001 PROVENANCE IN LLM REASONING DISTILLATION 002

003 **Anonymous authors**

004 Paper under double-blind review

## 005 ABSTRACT

006 Reasoning distillation, a cost-effective approach for enhancing student model per-  
007 formance, has attracted increasing attention. It typically leverages a large teacher  
008 model to generate reasoning paths, which are then used to fine-tune a student  
009 model so that it mimics the teacher’s behavior in training contexts. However, pre-  
010 vious approaches have lacked a detailed analysis of the origins of the distilled  
011 model’s capabilities. It remains unclear whether the student can maintain consis-  
012 tent behaviors with the teacher in novel test-time contexts, or whether it regresses  
013 to its original output patterns, raising concerns about the generalization of dis-  
014 tillation models. To analyse this question, we introduce a cross-model Reason-  
015 ing Distillation Provenance Tracing framework. For each action (e.g., a sentence)  
016 produced by the distilled model, we obtain the predictive probabilities assigned by  
017 the teacher, the original student, and the distilled model under the same context.  
018 By comparing these probabilities, we classify each action into four categories: (i)  
019 teacher-originated actions, (ii) student-originated actions, (iii) pre-existing actions  
020 in both models not enhanced by distillation, and (iv) pre-existing actions boosted  
021 through distillation. By systematically disentangling the provenance of each ac-  
022 tion, we experimentally demonstrate that, in test-time contexts, the distilled model  
023 can indeed generate teacher-originated actions, which correlate with and plausibly  
024 explain observed performance on distilled model. Building on this analysis, we  
025 further propose a teacher-guided data selection method. Unlike prior approach that  
026 rely on heuristics (e.g., selecting data most aligned with the student’s original dis-  
027 tribution), our method directly compares teacher–student divergences on the train-  
028 ing data, providing a principled selection criterion. We validate the effectiveness  
029 of our approach across multiple representative teacher models (Deepseek, QwQ,  
030 GPT-OSS-120B) and diverse student models (Qwen2.5-7B-Instruct, Qwen4-4B-  
031 Base, Qwen3-8B-Base, Qwen3-4B-Instruct-2507). The results highlight the util-  
032 ity of our provenance-tracing framework and underscore its promise for reasoning  
033 distillation. We hope to share Reasoning Distillation Provenance Tracing, along  
034 with our insights into reasoning distillation, with the community.

## 035 1 INTRODUCTION

036 The rapid development of reinforcement learning (RL) techniques (Schulman et al., 2017; Shao  
037 et al., 2024; Ahmadian et al., 2024) and resulting large-scale reasoning models (DeepSeek-AI, 2025;  
038 OpenAI, 2025; Team, 2025b) has accelerated the growth of distillation researches, especially reasoning  
039 distillation (Zhao et al., 2025; Guha et al., 2025). Early work on reasoning distillation focused  
040 on creating high-quality open-source datasets (Zhao et al., 2025; Guha et al., 2025; NVIDIA, 2025),  
041 and recent studies have focused on curating and filtering distillation samples to improve efficiency  
042 and performance (Zhang et al., 2025; Li et al., 2025). However, these approaches primarily focus on  
043 model performance and fail to provide an explanatory analysis of the sources of the model’s output  
044 in test-time contexts. As a result, it remains unclear whether the student model has successfully  
045 inherited the knowledge and reasoning logic from the teacher model, raising concerns about the  
046 generalization of distillation models (Hinton et al., 2015b).

047 Concretely, as illustrated in Figure 1, reasoning distillation involves two main steps. First, specific  
048 contexts (e.g., high-quality questions, or high-quality questions augmented with partially generated  
049 answers) are provided to the teacher model, which then produces the next action via sampling.

Second, these context–action pairs are used to train the student model, encouraging it to reproduce the teacher’s actions under the same contexts. At test time, however, the distilled student faces new contexts. It remains unclear whether the student will continue to follow the teacher’s behavior instead of falling back on its original output distribution.

To analyse the above problem, we propose a cross-model Reasoning Distillation Provenance Tracing framework. Concretely, we evaluate three open-source distilled models: DeepSeek-Distill-Qwen-7B (DeepSeek-AI, 2025), DeepSeek-R1-0528-Qwen3-8B (DeepSeek-AI, 2025), and LIMO-v2 model (Ye et al., 2025). To analyze the patterns in distilled models’ outputs under test contexts, we collect multiple responses from them on GPQA-D (Rein et al., 2024) and AIME24. Each response is re-input into the distilled model, the corresponding teacher model, and the original student model. This procedure allows us to obtain, under the same test scenario, the probability assigned by each model to the next action produced by the distilled model. By comparing these probabilities, we can naturally trace the provenance of every action. For example, if the teacher model assigns significantly higher probability to an action than the original student model does, we can attribute that action to teacher-originated actions, since the distilled model’s ability to produce it mainly derives from teacher model’s knowledge transferred during distillation. For further details about other action types, see Section 3.

To this end, we observe that distilled models can reproduce teacher-originated actions in new test contexts. Moreover, these actions are correlated with correct responses on the test set, which helps explain the generalization gains achieved through distillation. This observation further motivates our training design: we hypothesize that when the training data contains a higher proportion of teacher-originated actions, the distilled model attains better final performance. To validate this hypothesis, we propose a teacher-guided data selection strategy that compares the probabilities assigned by the teacher and student models on the training data and prioritizes examples that contain more teacher-originated actions. We then evaluate its effectiveness across multiple settings, including representative teacher models from different families (DeepSeek-R1-671B (DeepSeek-AI, 2025), QwQ-32B (Team, 2025b), GPT-OSS-120B (OpenAI, 2025)) and various student models (Qwen2.5-7B-Instruct, Qwen4-4B-Base, Qwen3-8B-Base, Qwen3-4B-Instruct-2507).

Our contributions can be summarised as follows: (1) We propose Reasoning Distillation Provenance Tracing, a systematic method to disentangle the origins of each action, through fine-grained categorization into teacher-originated, student-originated, shared, and boosted actions. This offers a principled approach for analyzing whether distillation genuinely transfers reasoning ability or merely reinforces pre-existing patterns. (2) Through analysis on reasoning benchmarks, we show that distilled models can generate teacher-originated actions even in unseen test scenarios. These actions are correlated with correctness, offering a quantitative explanation for why reasoning distillation improves generalization. (3) Building on the provenance analysis, we introduce a teacher-guided data selection strategy that prioritizes training samples rich in teacher-originated actions. Unlike heuristic method, our approach leverages explicit teacher–student divergence as a selection criterion. Experiments across diverse teacher–student pairs demonstrating performance gains in our settings.

## 2 RELATED WORK

### 2.1 REASONING DISTILLATION

Distilling the reasoning abilities of large reasoning models has been an important problem since their emergence (Kim & Rush, 2016; Ho et al., 2023; Ko et al., 2024; 2025). DeepSeek (DeepSeek-AI, 2025) pioneered this line of work by showing that supervised fine-tuning on the outputs of

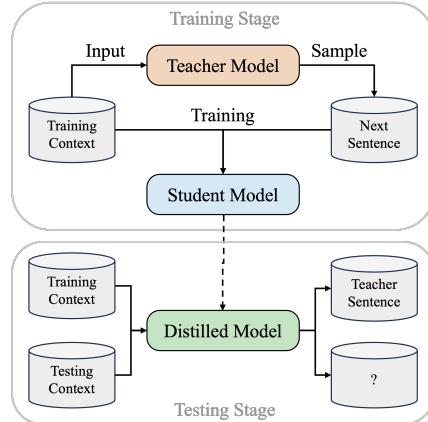


Figure 1: Motivation. In reasoning distillation, the student model learns in the training stage to produce actions consistent with the teacher model within the training context. However, at test time, it remains unclear whether the distilled model will continue to output actions aligned with the teacher model instead of degrading to the outputs of the student model, raising concerns about the generalization of distillation models (Hinton et al., 2015b).

108 a reasoning teacher, which is also the approach we focus on in this paper, can dramatically enhance the reasoning abilities of smaller models. Numerous subsequent projects (e.g., HuggingFace  
 109 OpenR1 (Hugging Face, 2025), OpenThoughts (Guha et al., 2025), a-m-team (Zhao et al., 2025),  
 110 NVIDIA AceReason (Liu et al., 2025b), Alibaba OmniThought (Guha et al., 2025), LIMO (Ye  
 111 et al., 2025), Tencent DeepMath (He et al., 2025)) have devoted substantial effort to constructing  
 112 large-scale corpora of challenging reasoning problems paired with teacher responses, using rigorous  
 113 quality filtering, correctness checks, and diversity-aware curation. Most recently, GRAPE  
 114 (Zhang et al., 2025) preferentially selects examples whose likelihoods best match the student’s cur-  
 115 rent distribution, thereby steering training toward data that is already well aligned with the student.  
 116 **Rather than focusing solely on artificially designed rules and heuristic rules, we aim to quan-  
 117 tify the sources of a distilled model’s capabilities and introduce a data selection criterion that  
 118 focuses on sentences whose probabilities indicate stronger teacher-originated behavior. This  
 119 provenance-aware criterion complements prior student-only selection and provides an explicit  
 120 cross-model signal for reasoning transfer.**  
 121

## 2.2 MODEL AUDITING

122 Another closely related area is model auditing, a growing line of work that studies (Song &  
 123 Shmatikov, 2019; Carlini et al., 2022; Dekoninck et al., 2024) auditing generative models to under-  
 124 stand what data they memorize and to attribute outputs back to underlying data sources. **In contrast,  
 125 our work targets model-level provenance in a distillation setting: rather than asking whether  
 126 specific data are memorized, we aim to trace which upstream models are the sources of a given  
 127 output, shifting the focus from data membership to the lineage of the models themselves.**  
 128

129 Due to page limitations, we provide further discussions on related work in the Appendix A.6.

## 3 REASONING DISTILLATION PROVENANCE TRACING

130 In this section, we first provide necessary notations in Subsection 3.1 and introduce Reasoning  
 131 Distillation Provenance Tracing in Subsection 3.2. We then apply Reasoning Distillation Provenance  
 132 Tracing to three widely used open-source models (Deepseek-Distill-Qwen-7B, DeepSeek-R1-0528-  
 133 Qwen3-8B and LIMO-v2 model) and present the results of the analysis in Subsection 3.3.  
 134

### 3.1 NOTATIONS

135 Specifically, let  $M_T$  denote the teacher model,  $M_S$  the student model, and  $\{Q_{train}^i\}$  a set of high-  
 136 quality training questions, where  $Q_{train}^i$  denotes the  $i$ -th question. Reasoning distillation first sam-  
 137 ples responses from  $M_T$  on  $\{Q_{train}^i\}$ , yielding the training set  $\mathcal{D}_{train} = \{(Q_{train}^i, \tau^i)\}$ , where  
 138  $\tau_i = \{a_{(i,j)} | j = 1, \dots, L_i\}$  denotes the trajectory for the  $i$ -th question, consisting of  $L_i$  actions gen-  
 139 erated by  $M_T$ . The student model  $M_S$  is then trained by minimizing the cross-entropy loss between  
 140 its predicted next-token-actions and the teacher’s actions in trajectory  $\tau_i$  under the same context  
 141 (e.g., input  $(Q_{train}^i, a_{(i,1)}, a_{(i,2)})$ , output  $a_{(i,3)}$ ). This ensures that distilled model  $M_D$  produces  
 142 reasoning trajectories mimicking to those of  $M_T$  when presented with the same context.  
 143

### 3.2 METHOD

144 As stated in the Section 1, our goal is to analyze whether the distilled model  $M_D$  can still produce  
 145 actions similar to those of the teacher model  $M_T$  in new contexts. The most direct approach is to  
 146 input the same context into both models ( $M_T$  and  $M_D$ ) and compare their actions. Yet this approach  
 147 faces two challenges. First, reasoning outputs are typically long, and even when segmented step by  
 148 step, the number of sentences remains large. Iteratively truncating at each sentence boundary and  
 149 re-input both models to generate new actions is prohibitively expensive. Second, it is difficult to  
 150 accurately evaluate the similarity between the newly generated actions of the two models.  
 151

152 To address these issues, we shift perspective and instead sample exclusively from the distilled model  
 153  $M_D$  on the test set  $\{Q_{test}^i\}$ . We first obtain the trajectory on test set  $\mathcal{D}_{test} = \{(Q_{test}^i, \tau^i)\}$ . Sub-  
 154 sequently, we feed  $\tau^i$  back into the three models ( $M_T$ ,  $M_S$ , and  $M_D$ ) and analyse each component  
 155 action  $a_{(i,j)}$ . For each action, since token-level comparison is sometimes difficult due to possible  
 156 vocabulary mismatches between  $M_T$  and  $M_D$ . We use a coarser alternative, comparing at the sen-  
 157 tence level. In this way, we define the probability of producing  $a_{(i,j)}$  as the geometric mean of  
 158 per-token probabilities  $p_{(i,j)} = \exp(\text{mean}(\log(p_k)))$ , where  $p_k$  denote the probability of  $k$ -th token  
 159 contained with the sentence  $a_{(i,j)}$ .  
 160

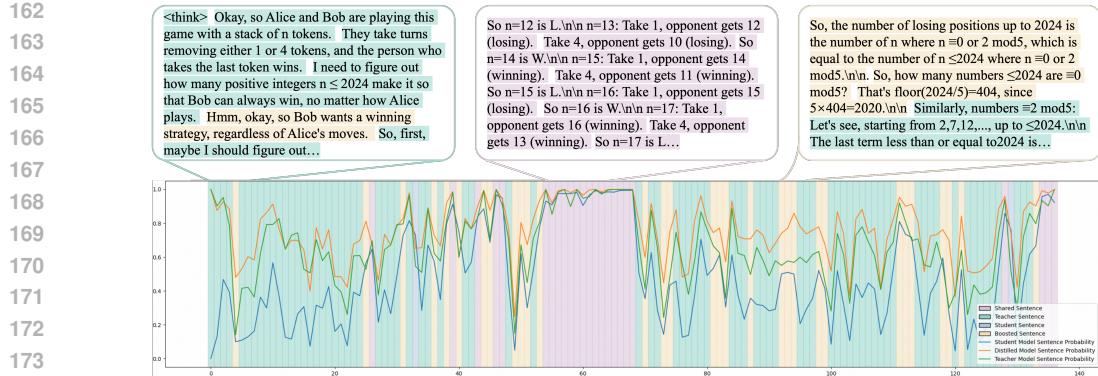


Figure 2: An illustration of analysis using Reasoning Distillation Provenance Tracing. The horizontal axis denotes the action position (i.e., the sentence order), while the vertical axis shows the probability assigned by different models (indicated by colored curves in the foreground) in producing that action under the same context. The background colors indicate the action types. Some example of different action types are also shown at the top for illustration, where the blank spaces indicate the segmentation boundaries of the response.

For each  $a_{(i,j)}$ , we can obtain three output probabilities under the same context:  $p_{(i,j)}^T$  from  $M_T$ ,  $p_{(i,j)}^S$  from  $M_S$ , and  $p_{(i,j)}^D$  from  $M_D$ . As shown in Figure 2, the blue line corresponds to  $p_{(i,j)}^S$ , the orange line to  $p_{(i,j)}^D$ , and the green line to  $p_{(i,j)}^T$ .

Overall, since the trajectory  $\tau^i$  is sampled from the distilled model  $M_D$ , the orange curve ( $p_{(i,j)}^D$ ) tends to be the largest on average. Furthermore, four distinct patterns can be observed. Specifically:

**Pre-existing actions in both models not enhanced by distillation** (hereafter referred to as Shared Sentence): This includes actions such as those between the 54th and 68th steps, where the output probabilities of all models are nearly identical. These actions are originally present in both the teacher and the student models, and distillation does not further increase their probabilities.

**Pre-existing actions boosted through distillation** (hereafter referred to as Boosted Sentence): Similar to the first type,  $p_{(i,j)}^T$  and  $p_{(i,j)}^S$  remain close, but  $p_{(i,j)}^D$  differs significantly (and is typically higher in practice, since trajectories are sampled from  $M_D$ ). These actions also exist in both the teacher and student models prior to distillation, but their probabilities are amplified through training with distilled data.

**Student-originated actions** (hereafter referred to as Student Sentence) and **teacher-originated actions** (hereafter referred to as Teacher Sentence): When there is a large discrepancy between  $p_{(i,j)}^S$  and  $p_{(i,j)}^T$ , the distilled model  $M_D$  still outputs the action, suggesting the action is more consistent with the model assigning higher likelihood. Note that a Teacher Sentence does not imply that the action is entirely absent from the student model, but rather that it is primarily originated from the teacher. The same applies to a Student Sentence.

### 3.3 ANALYSIS ON OPEN-SOURCE MODELS

In this subsection, we apply Reasoning Distillation Provenance Tracing to DeepSeek-Distill-Qwen-7B (Ds-7B), DeepSeek-R1-0528-Qwen3-8B (Ds-8B) and LIMO-v2 model to analyze the source of each sentence produced by the distilled models in new testing context. Specifically, we sample the open-source models on two commonly used reasoning benchmarks: AIME24 and GPQA-D. AIME24 uses 16 completions per question, and GPQA-D uses 8.

For each completed response, we segment the output according to the following rules: (1) Special tokens (e.g., <think>) are treated as individual actions, as they carry specific semantics and represent important behavior. (2) For the remaining text, we split sentences using the pattern: punctuation + optional whitespace + uppercase letter, corresponding to `re.compile(r'([.?!]\s*)\s*([A-Z])')`. This segmentation procedure covers most common English sentence structures.

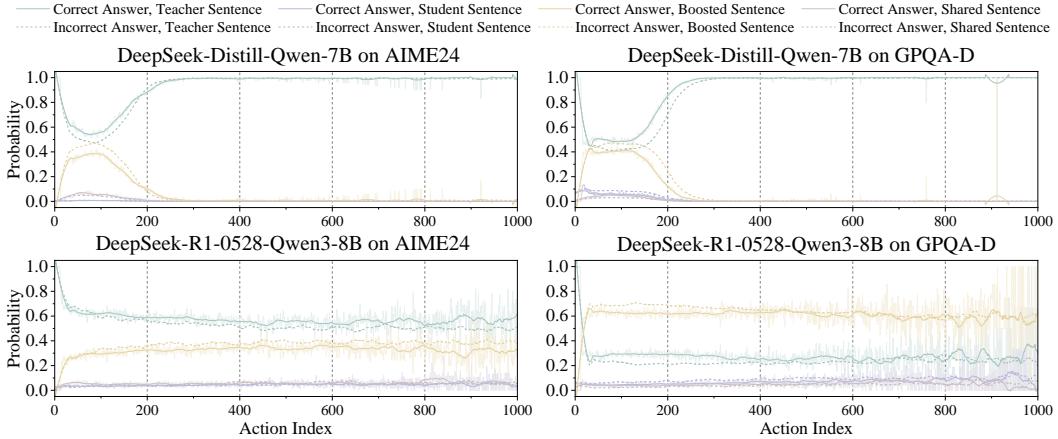


Figure 3: **Analysis results on open-source models.** We apply Reasoning Distillation Provenance Tracing to Deepseek-Distill-Qwen-7B and DeepSeek-R1-0528-Qwen3-8B, analyzing the probability of producing different actions at each action position on two reasoning benchmarks (AIME24 and GPQA-D). The x-axis denotes the action index, and the y-axis shows the probability of producing that action.

Then, we feed the sampled trajectories back into the corresponding teacher and student models to obtain the output probability for each token, and consequently (by taken geometric average of token probability), the probability for each action ( $p_{(i,j)}^S$ ,  $p_{(i,j)}^D$  and  $p_{(i,j)}^T$ ). We then define  $\Delta_{SD} = p_{(i,j)}^S - p_{(i,j)}^D$ ,  $\Delta_{TD} = p_{(i,j)}^T - p_{(i,j)}^D$  and  $\Delta_{TS} = p_{(i,j)}^T - p_{(i,j)}^S$ . Using two thresholds,  $\alpha$  and  $\beta$ , we can perform the tracing provenance:

$$\begin{cases} \text{Shared Sentence,} & \text{if}(|\Delta_{SD}| \leq \alpha \wedge |\Delta_{TD}| \leq \alpha \wedge |\Delta_{TS}| \leq \alpha), \\ \text{Teacher Sentence,} & \text{else if}(\Delta_{TS} > \beta), \\ \text{Student Sentence,} & \text{else if}(-\Delta_{TS} > \beta), \\ \text{Boosted Sentence,} & \text{else if}(|\Delta_{TS}| < \beta). \end{cases}$$

where  $p_{(i,j)}^D$ ,  $p_{(i,j)}^S$  and  $p_{(i,j)}^T$  denote the probabilities assigned to action  $a_{(i,j)}$  by the distilled model  $M_D$ , the student model  $M_S$ , and the teacher model  $M_T$ , respectively. We follow the evaluation order: Shared  $\rightarrow$  Teacher  $\rightarrow$  Student  $\rightarrow$  Boosted. The role of  $\alpha$  is to filter out relatively small probability differences (such as those between sentences 55 and 65 in Figure 2) to prevent them from influencing the analysis.  $\beta$  helps to more clearly differentiate between various action types, and its value can be determined adaptively. Due to page limitations, the selection of  $\alpha$  and  $\beta$  is detailed in Appendix A.2.

Based on this classification scheme, we compute, for each action position, the proportion of each action type and interpret this proportion as the probability of outputting that action type at the given position. For example, for all answers at the third action position (i.e., the third sentence), we calculate the proportion of actions belonging to Teacher Sentence and treat it as the probability that the model outputs Teacher Sentence at that position. **It is worth noting that the number of actions varies across different answers; therefore, we focus primarily on the earlier action positions, where sufficient data is available to yield reliable statistics.** In addition, we compute the average number of tokens for all actions at each position, and, at 4k-token intervals, mark in the action positions corresponding to the average number of tokens required to reach that position. The results are shown in Figure 3 and Figure 4.

We observe three phenomena (see Appendix A.3 for more analyses) that help explain the observed benefits of reasoning distillation in novel test-time settings:

#### (1) Higher Teacher Sentence probability in the early inference stage.

As shown in Figure 3 and Figure 4, the light-green line (—) exhibits larger values in the early action index, meaning that the probability of outputting a Teacher Sentence is relatively high in the early stages of inference. We attribute this to two factors. First, the student model lacks the ability to properly generate the token <think>, and subsequent steps may be influenced by the

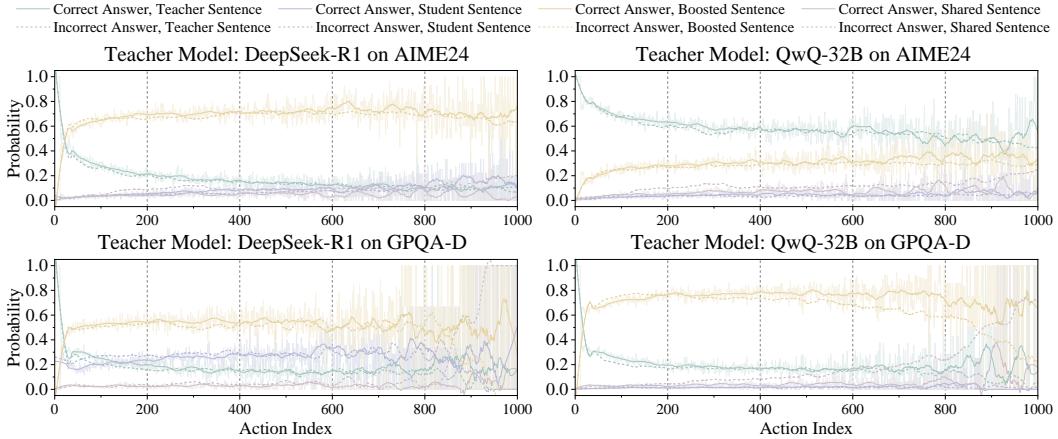


Figure 4: Analysis results on LIMO-v2 model (Ye et al., 2025). We apply Reasoning Distillation Provenance Tracing to LIMO-v2 model, analyzing the probability of producing different actions at each action position on two reasoning benchmarks (AIME24 and GPQA-D). Since LIMO-v2 model uses multiple teacher models, we tried different teacher models on all samples. The x-axis denotes the action index, and the y-axis shows the probability of producing that action.

first action. However, this effect is limited, as evidenced by the rapid increase in the probability of Boosted Sentence (i.e., the light-yellow line (—) rises sharply in the early phase.). Figure 2 and Figure 11 further illustrate this point, showing that other actions typically emerge after only a few steps. Second, we observe that many early actions focus on analyzing the input and planning subsequent steps. We hypothesize that forcing such behavior early on may be a unique pattern of the teacher model, which explains why Teacher Sentences are more likely to appear at the beginning of inference.

#### (2) Student-internal patterns are often activated but not uniformly beneficial during reasoning distillation.

In LIMO-v2 paper, the authors found that if a model already contains sufficient reasoning knowledge, reasoning distillation can activate it with minimal data scale and yield strong performance. In this work, we find that reasoning distillation not only activates reasoning knowledge, but also triggers other latent patterns embedded within the Boosted Sentence. As shown in Figure 3 and Figure 4, we observe that the mean of the pointwise sum of the light-green (—) and light-yellow (—) lines exceeds 0.7, indicating that at each position the predicted action is likely to be one of these two action types. This means that, in distilled models, the majority of output action types are either Teacher Sentences or Boosted Sentences. More Teacher Sentences originate from the student model learning directly from the teacher model. And more Boosted Sentences suggests that reasoning distillation consistently activates student-internal patterns.

To analyze how reasoning distillation helps in novel contexts, we further conduct a quantitative analysis of action-type probabilities in correct versus incorrect responses. As shown in Figure 4, on LIMO-v2 model, Boosted Sentences emerge in later stages with consistently high probability in correct responses across different teacher models (i.e., in later stages, the light-yellow solid line (—) remains above the light-yellow dashed line (—)). In the early stages, correct responses are more likely to rely on Teacher Sentences (i.e., the light-green solid line (—) stays above the light-green dashed line (—)). However, Figure 3 also reveals the opposite trend in smaller models: Boosted Sentences consistently appear with higher probability in incorrect responses. This suggests that not all student-internal patterns are worth activating. The authors of LIMO-v2 paper also find that small models (e.g., 7B) trained with the same method fail to achieve good performance. Taken together with our findings, one concludes that reasoning distillation often activates internal patterns in student model, but not all of these activations are beneficial.

#### (3) Teacher sentences are highly correlated with performance gains.

As shown in Figure 3, across different models (DeepSeek-R1-0528-Qwen3-8B and DeepSeek-Distill-Qwen-7B) and test sets (AIME24 and GPQA-D), Teacher Sentence tends to be assigned higher probabilities to correct answers (i.e., the light-green solid line (—) stays above the light-green dashed line (—)). This result is also intuitive: since the teacher model achieves higher

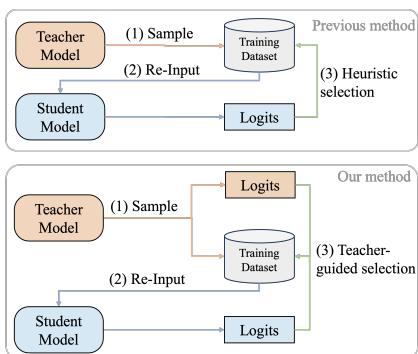
324 performance on the test sets, the student model benefits from producing outputs more aligned with  
 325 the teacher, thereby increasing its likelihood of answering correctly.  
 326 It remains an open question whether Teacher Sentence is equally beneficial for larger models, and we  
 327 believe further experimentation is needed. Although Figure 4 clearly shows that Teacher Sentence  
 328 assigns higher probabilities to correct answers in the early stages (while probabilities for correct and  
 329 incorrect answers converge in later stages), the evidence is limited. LIMO-v2 dataset contains only  
 330 800 samples, designed primarily to activate student-internal patterns and validate conclusions in the  
 331 original paper. Thus, it is unclear whether, at larger training scales, Teacher Sentence would still  
 332 yield significantly higher probabilities for correct answers in the mid-to-late stages. Due to resource  
 333 constraints, this work focuses on models of 8B parameters or fewer. We hope to validate our findings  
 334 on larger models in future work.

335 **In summary, our Reasoning Distillation Provenance Tracing shows that in new test settings,**  
 336 **distilled models not only produce sentences originating from the teacher, but these Teacher**  
 337 **Sentence are especially prominent in early reasoning phases and associated with answer cor-**  
 338 **rectness. At the same time, reasoning distillation can also activate patterns already latent**  
 339 **within the student (Boosted Sentences), whose contributions vary by model size and reasoning**  
 340 **phases, and are not uniformly beneficial.**

## 341 4 BEYOND EXPLANATION, GUIDING TRAINING IN REVERSE: 342 TEACHER-GUIDED DATA SELECTION

344 Building on the analysis in Section 3, a natural training insight is: **if we can increase the proportion**  
 345 **of Teacher Sentence in the training data and explicitly leverage teacher-student divergences**  
 346 **to select examples, we may front-load and amplify the observed generalization benefits.** There-  
 347 **fore, in this section, we first propose a teacher-guided data selection method in Section 4.1. Then we**  
 348 **validate its effectiveness across multiple teacher-student pairs and reasoning benchmarks in Section**  
 349 **4.2.**

### 350 4.1 TEACHER-GUIDED DATA SELECTION



362 Figure 5: Comparison between teacher-  
 363 guided data selection and the previous  
 364 method (Zhang et al., 2025).

365 Specifically, for each question with multiple responses from **the same** teacher model, we count the number  
 366 of Teacher Sentences in each response and prioritize for training the response with the largest count.

368 As shown in Figure 5, the previous method (Zhang et al., 2025) feeds the training data into the  
 369 student model before training, computes heuristics metrics based on the logits of student model, and  
 370 selects samples that more aligned with the student model’s original distribution. In contrast, **our**  
 371 **proposed method provides a clearer objective for data selection: prioritizing samples where**  
 372 **the teacher and student models differ the most.**

373 Moreover, our method introduces only an acceptable additional cost in practice. (1) When one  
 374 wishes to sample responses, similar to the previous method (Zhang et al., 2025), our approach  
 375 incurs no extra overhead, since logits can be obtained directly during sampling. (2) When one  
 376 wishes to leverage existing open-source distillation datasets for filtering, some additional cost is  
 377 required. However, it is important to note that re-feeding the generated sequences into the model to  
 extract logits only requires a single forward pass, which is significantly faster than token-by-token  
 generation. Given sufficient GPU memory, we consider the additional time cost to be acceptable.

378 

## 4.2 EXPERIMENTS

380 We aim to evaluate the effectiveness of teacher-guided data selection across diverse training settings.  
 381 To this end, we conduct experiments with three distinct teacher models (Deepseek-R1-671B, QwQ-  
 382 32B, and GPT-OSS-120B) and four student models: Qwen3-4B-Base, Qwen3-8B-Base, Qwen2.5-  
 383 7B-Instruct, and Qwen3-4B-Instruct-2507. We use two high-quality reasoning datasets, AceReason-  
 384 1.1-SFT (Liu et al., 2025b) and OpenThought3-1.2M (Guha et al., 2025). Each question in these  
 385 datasets is paired with multiple candidate answers sampled from the teacher model. Further de-  
 386 tails are provided in Appendix A.5. For each question, we then select a single response using one  
 387 of three strategies: (i) random selection, which performs similarly to vanilla reasoning distillation  
 388 (“Vanilla”); (ii) GRAPE (Zhang et al., 2025) (“GRAPE”); and (iii) our method (“Ours”). All strate-  
 389 gies yield training sets of identical size.

390 For evaluation, we select four widely adopted benchmarks known for their challenging reasoning  
 391 demands: AIME24, AIME25, MATH500 (Lightman et al., 2023), and OlympiadBench (He et al.,  
 392 2024). We report the accuracy results.

393 

### 4.2.1 MAIN EXPERIMENTS

394 As shown in Tables 1, our method achieves the best performance. On average, it improves results  
 395 by 1.7%–2.5%.

396 Table 1: Comparison with different data selection strategy. Each experimental setting is denoted as  
 397 Teacher Model + Student Model + Data Source.

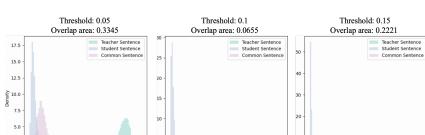
	AIME24	AIME25	MATH500	OlympiadBench	Average
<i>Deepseek-R1 + Qwen3-4B-Base + AceReason-1.1-SFT</i>					
Vanilla	44.4	33.3	<b>91.2</b>	55.9	56.2
GRAPE (Zhang et al., 2025)	43.0	34.1	88.6	54.8	55.1
Ours	<b>49.3</b> <sub>+4.9</sub>	<b>37.9</b> <sub>+4.6</sub>	90.8 <sub>-0.4</sub>	<b>56.6</b> <sub>+0.7</sub>	<b>58.7</b> <sub>+2.5</sub>
<i>Deepseek-R1 + Qwen3-8B-Base + AceReason-1.1-SFT</i>					
Vanilla	55.1	39.9	91.2	57.5	60.9
GRAPE (Zhang et al., 2025)	54.2	38.1	91.8	58.4	60.6
Ours	<b>57.3</b> <sub>+2.2</sub>	<b>41.5</b> <sub>+1.6</sub>	<b>92.8</b> <sub>+1.6</sub>	<b>58.7</b> <sub>+1.2</sub>	<b>62.6</b> <sub>+1.7</sub>
<i>QwQ-32B + Qwen2.5-7B-Instruct + OpenThought3-1.2M</i>					
Vanilla	43.5	35.6	89.8	55.1	56.0
GRAPE (Zhang et al., 2025)	47.5	34.6	<b>91.4</b>	54.8	57.1
Ours	<b>48.1</b> <sub>+4.6</sub>	<b>36.3</b> <sub>+0.7</sub>	90.0 <sub>+0.2</sub>	<b>56.3</b> <sub>+1.2</sub>	<b>57.7</b> <sub>+1.7</sub>
<i>GPT-OSS-120B + Qwen3-4B-Instruct-2507 + AceReason-1.1-SFT</i>					
Vanilla	75.9	62.5	93.6	64.0	74.0
GRAPE (Zhang et al., 2025)	76.8	66.5	92.6	62.5	74.6
Ours	<b>77.9</b> <sub>+2.0</sub>	<b>68.3</b> <sub>+5.8</sub>	<b>94.6</b> <sub>+1.0</sub>	<b>64.9</b> <sub>+0.9</sub>	<b>76.4</b> <sub>+2.4</sub>

413 Table 2: Influence of different selection metrics.

	AIME24	AIME25	MATH500	OlympiadBench	Average
Maximize Absolute Count	49.3	37.9	90.8	56.6	58.7
Longest	48.1	37.5	90.0	55.9	57.9
Relative Proportion	46.9	35.0	88.8	55.1	56.4
Vanilla	44.4	33.3	91.2	55.9	56.2
Minimize Absolute Count	42.9	35.2	87.8	54.2	55.0

420 

### 4.2.2 ABLATION EXPERIMENTS



421 Figure 6: Illustration of  $\beta$  selection for  
 422 the first training setting.

423 **(1) How to Leverage Teacher Sentences for Data Se-  
 424 lection?** We generally aim to include more Teacher Sen-  
 425 tences in training data, where “more” can be interpreted  
 426 in two ways: by the absolute count of Teacher Sentences  
 427 within a response (“Maximize Absolute Count”) or by  
 428 their relative proportion (“Relative Proportion”). Figure  
 429 3 shows that, across most action positions in correct re-  
 430 sponses, Teacher Sentences receive higher output proba-  
 431 bilities from the distilled model. This observation sug-  
 432 gests two principles for effective data selection: (i) choose responses that contain as many reasoning  
 433 actions as possible (“Longest”) so that the training signal can influence the maximum number of  
 434 action positions, and (ii) increase the absolute number of those actions that correspond to Teacher

432 Sentences (“Maximize Absolute Count”) to raise the model’s probability of producing them at those  
 433 positions.

434 To compare these metrics, we conduct an ablation study using Deepseek-R1 as the teacher, Qwen3-  
 435 4B-Base as the student, and AceReason-1.1-SFT as the data source. The results in Table 2 show that  
 436 **Maximize Absolute Count yields the best performance, whereas Minimize Absolute Count**  
 437 **performs the worst, indirectly corroborating our hypothesis.**

439 Table 3: Impact of different  $\beta$  values on accuracy.

	AIME24	AIME25	MATH500	OlympiadBench	Average
$\beta = 0.05$	47.9	37.1	89.8	57.0	58.0
$\beta = 0.1$	49.3	37.9	90.8	56.6	58.7
$\beta = 0.15$	46.3	37.5	90.2	56.3	57.6

440  
 441  
 442  
 443  
 444 **(2) How to determine  $\beta$ ?** In Section 3, we describe how to set  $\beta$  at test time. Here, we assess  
 445 whether the same procedure is effective on the training data. We conduct validation experiments  
 446 using DeepSeek-R1 as the teacher model, Qwen3-4B-Base as the student model, and AceReason-  
 447 1.1-SFT as the training data source. As shown in Figure 6, before training we observe that  $\beta = 0.1$   
 448 produces the cleanest partition, with minimal overlap across action types. As shown in Table 3, the  
 449 optimal  $\beta$  determined before training achieved the best results in the ablation experiment. Moreover,  
 450 other near-optimal  $\beta$  also outperform the Vanilla baseline, indicating that performance is not highly  
 451 sensitive to the choice of  $\beta$ . **Together, the post-training results, along with the pre-training**  
 452 **analysis, consistently support our data-driven  $\beta$ -selection strategy as a principled, training-**  
 453 **free, and effective approach.** For further discussion, see Appendix A.2.

454 **(3) Effects in different domain.** To evaluate the ef-  
 455 fectiveness of our method in different domain, we used  
 456 GPT-OSS-120B to sample 10k questions from Open-  
 457 ScienceReasoning\_2 (NVIDIA Corporation, 2024), gen-  
 458 erating nine candidate responses per question. We then  
 459 selected a single response for training using different se-  
 460 lection strategies. Results are reported in Table 4. We find  
 461 that our method is effective in the scientific domain and,  
 462 compared to GRAPE, exhibits stronger generalization in  
 463 math domain. Analysis on the test set further indicates  
 464 that the Teacher Sentence in science shares commonali-  
 465 ties with that in mathematics, which we hypothesize helps  
 466 explain why training on scientific data can improve per-  
 467 formance on the math test set. Additional analyses are provided in Appendix A.3.3.

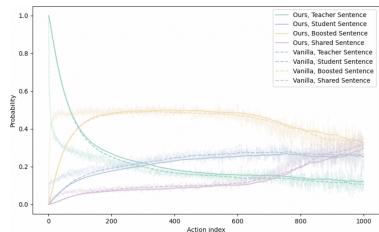
468 Table 4: Comparison with different data selection strategy on scientific domain.

	GPQA-D	AIME24	AIME25	MATH500	OlympiadBench
<i>GPT-OSS-120B + Qwen3-4B-Base + OpenScienceReasoning_2</i>					
Vanilla	39.9	24.2	24.6	85.8	44.0
GRAPE (Zhang et al., 2025)	40.9	24.6	22.1	84.2	44.3
Ours	<b>42.9<sub>+3.0</sub></b>	<b>25.2<sub>+1.0</sub></b>	<b>24.8<sub>+0.2</sub></b>	<b>87.4<sub>+1.6</sub></b>	<b>48.0<sub>+4.0</sub></b>

473 **(4) Does the distilled model increase the probability of outputting the Teacher Sentence?** We  
 474 apply the Reasoning Distillation Provenance Tracing framework to our trained models. The re-  
 475 sults are shown in Figure 7. Compared with the Vanilla baseline, within the range where estimates  
 476 are statistically reliable (e.g., the first 800 action indices), our method increases the probability of  
 477 outputting the Teacher Sentence and correspondingly decreases that of the Student Sentence.

## 478 CONCLUSION

479 In this work, we address a fundamental yet under-explored question in reasoning distillation: does  
 480 the distilled model truly inherit the teacher’s pattern, or does it revert to its prior patterns when  
 481 confronted with new contexts? To answer this, we introduce Reasoning Distillation Provenance Tracing,  
 482 observe phenomena and quantify the evidences that help explain the observed benefits of reasoning  
 483 distillation in novel test-time settings. Building on these insights, we propose a teacher-guided data-  
 484 selection strategy and demonstrate its effectiveness on multiple settings. We hope our provenance-  
 485 tracing framework will inspire future research on cross-model behavior analysis, domain-aware data  
 selection, and more reliable distillation protocols for complex reasoning tasks.



479 Figure 7: **Analysis results of the**  
 480 **first training setting (Deepseek-R1 +**  
 481 **Qwen3-4B-Base + AceReason-1.1-**  
 482 **SFT) on AIME24.**

486 REFERENCES  
487

488 Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin,  
489 Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce style optimization for learn-  
490 ing from human feedback in llms. *arXiv preprint arXiv:2402.14740*, 2024.

491 Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan  
492 Zhang. Quantifying memorization across neural language models. In *The Eleventh International  
493 Conference on Learning Representations*, 2022.

494 Runjin Chen, Zhenyu Zhang, Junyuan Hong, Souvik Kundu, and Zhangyang Wang. Seal: Steerable  
495 reasoning calibration of large language models for free. *arXiv preprint arXiv:2504.07986*, 2025.

496 DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning,  
497 2025. URL <https://arxiv.org/abs/2501.12948>.

498 Jasper Dekoninck, Mark Niklas Müller, Maximilian Baader, Marc Fischer, and Martin Vechev.  
499 Evading data contamination detection for language models is (too) easy. *arXiv preprint  
arXiv:2402.02823*, 2024.

500 Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D Goodman. Cogni-  
501 tive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv  
502 preprint arXiv:2503.01307*, 2025.

503 Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A  
504 survey. *International journal of computer vision*, 129(6):1789–1819, 2021.

505 Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna  
506 Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer, Liangyu  
507 Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su,  
508 Wanja Zhao, John Yang, Shreyas Pimpalgaonkar, Kartik Sharma, Charlie Cheng-Jie Ji, Yichuan  
509 Deng, Sarah Pratt, Vivek Ramanujan, Jon Saad-Falcon, Jeffrey Li, Achal Dave, Alon Albalak,  
510 Kushal Arora, Blake Wulfe, Chinmay Hegde, Greg Durrett, Sewoong Oh, Mohit Bansal, Saadia  
511 Gabriel, Aditya Grover, Kai-Wei Chang, Vaishaal Shankar, Aaron Gokaslan, Mike A. Merrill,  
512 Tatsunori Hashimoto, Yejin Choi, Jenia Jitsev, Reinhard Heckel, Maheswaran Sathiamoorthy,  
513 Alexandros G. Dimakis, and Ludwig Schmidt. Openthoughts: Data recipes for reasoning models,  
514 2025. URL <https://arxiv.org/abs/2506.04178>.

515 Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu,  
516 Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for  
517 promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint  
arXiv:2402.14008*, 2024.

518 Zhiwei He, Tian Liang, Jiahao Xu, Qiuwei Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu,  
519 Zhenwen Liang, Wenxuan Wang, Zhusong Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and  
520 Dong Yu. Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathemat-  
521 ical dataset for advancing reasoning. 2025. URL <https://arxiv.org/abs/2504.11456>.

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

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

526 Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. In  
527 *Proceedings of the 61st annual meeting of the association for computational linguistics (volume  
528 1: long papers)*, pp. 14852–14882, 2023.

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

540 Hugging Face. Open r1: A fully open reproduction of deepseek-r1, January 2025. URL <https://github.com/huggingface/open-r1>.  
 541  
 542

543 Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec  
 544 Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv*  
 545 *preprint arXiv:2412.16720*, 2024.

546 Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *Proceedings of the*  
 547 *2016 conference on empirical methods in natural language processing*, pp. 1317–1327, 2016.  
 548

549 Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined  
 550 distillation for large language models. *arXiv preprint arXiv:2402.03898*, 2024.  
 551

552 Jongwoo Ko, Tianyi Chen, Sungnyun Kim, Tianyu Ding, Luming Liang, Ilya Zharkov, and Se-  
 553 Young Yun. Distillm-2: A contrastive approach boosts the distillation of llms. *arXiv preprint*  
 554 *arXiv:2503.07067*, 2025.

555 Jinu Lee, Sagnik Mukherjee, Dilek Hakkani-Tur, and Julia Hockenmaier. Reasoningflow: Semantic  
 556 structure of complex reasoning traces. *arXiv preprint arXiv:2506.02532*, 2025.  
 557

558 Zhuang Li, Yuncheng Hua, Thuy Vu, Haolan Zhan, Lizhen Qu, and Gholamreza Haffari. Scar: Data  
 559 selection via style consistency-aware response ranking for efficient instruction-tuning of large lan-  
 560 guage models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational*  
*Linguistics (Volume 1: Long Papers)*, pp. 12756–12790, 2025.

561 Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan  
 562 Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. In *The Twelfth*  
*563 International Conference on Learning Representations*, 2023.  
 564

565 Kaiyuan Liu, Chen Shen, Zhanwei Zhang, Junjie Liu, Xiaosong Yuan, et al. Efficient reasoning  
 566 through suppression of self-affirmation reflections in large reasoning models. *arXiv preprint*  
 567 *arXiv:2506.12353*, 2025a.  
 568

569 Zihan Liu, Zhuolin Yang, Yang Chen, Chankyu Lee, Mohammad Shoeybi, Bryan Catanzaro, and  
 570 Wei Ping. Acereason-nemotron 1.1: Advancing math and code reasoning through sft and rl  
 571 synergy. *arXiv preprint arXiv:2506.13284*, 2025b.  
 572

573 NVIDIA. Nvidia nemotron nano 2: An accurate and efficient hybrid mamba-transformer reasoning  
 574 model, 2025. URL <https://arxiv.org/abs/2508.14444>.  
 575

576 NVIDIA Corporation. OpenScienceReasoning-2: A Dataset for Open Scientific Reason-  
 577 ing. <https://huggingface.co/datasets/nvidia/OpenScienceReasoning-2>,  
 578 September 2024. Hugging Face dataset.  
 579

580 OpenAI. gpt-oss-120b gpt-oss-20b model card, 2025. URL <https://arxiv.org/abs/2508.10925>.  
 581

582 David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Di-  
 583 rani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a bench-  
 584 mark. In *First Conference on Language Modeling*, 2024.  
 585

586 John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy  
 587 optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.  
 588

589 Romero A Ballas N Kahou SE. Fitnets: hints for thin deep nets. *Fitnets: hints for thin deep nets*,  
 590 2014.  
 591

592 Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang,  
 593 Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathemati-  
 594 cal reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.  
 595

596 Congzheng Song and Vitaly Shmatikov. Auditing data provenance in text-generation models. In  
 597 *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery &*  
*Data Mining*, pp. 196–206, 2019.

594 Qwen Team. Qwen3 technical report, 2025a. URL <https://arxiv.org/abs/2505.09388>.  
 595

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

598 Yiping Wang, Qing Yang, Zhiyuan Zeng, Liliang Ren, Liyuan Liu, Baolin Peng, Hao Cheng, Xuehai  
 599 He, Kuan Wang, Jianfeng Gao, et al. Reinforcement learning for reasoning in large language  
 600 models with one training example. *arXiv preprint arXiv:2504.20571*, 2025.

601

602 An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li,  
 603 Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin  
 604 Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang,  
 605 Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang,  
 606 Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan,  
 607 Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint*  
 608 *arXiv:2412.15115*, 2024.

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

611 Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the perfor-  
 612 mance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*,  
 613 2016.

614

615 Dylan Zhang, Qirun Dai, and Hao Peng. The best instruction-tuning data are those that fit. *arXiv*  
 616 *preprint arXiv:2502.04194*, 2025.

617 Han Zhao, Haotian Wang, Yiping Peng, Sitong Zhao, Xiaoyu Tian, Shuaiting Chen, Yunjie Ji, and  
 618 Xiangang Li. 1.4 million open-source distilled reasoning dataset to empower large language  
 619 model training, 2025. URL <https://arxiv.org/abs/2503.19633>.

620

621

622

623

624

625

626

627

628

629

630

631

632

633

634

635

636

637

638

639

640

641

642

643

644

645

646

647

648  
649

## A APPENDIX

650  
651

## A.1 FUTURE WORK

652  
653  
654  
655  
656

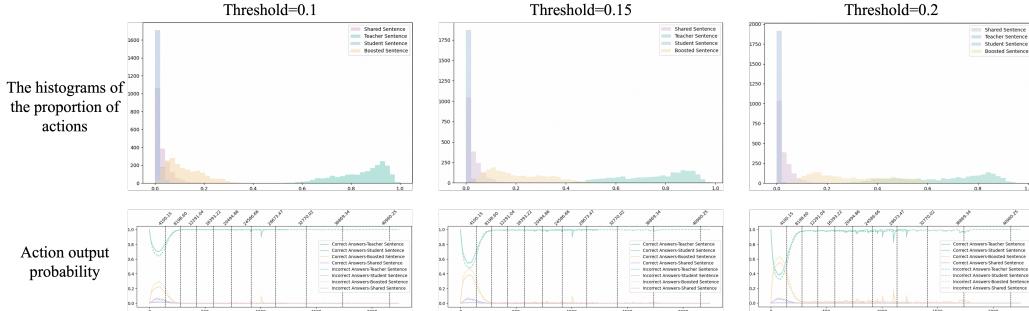
Future work will pursue three directions: (1) when more logits information is available, combine sequential probability matching methods to improve our data selection; (2) expand beyond the small-model regime studied here, imposed by resource constraints, to evaluate the effectiveness of our data-selection strategy on larger models; and (3) move past the single-teacher setting to investigate principled approaches to data selection when training with ensembles of diverse teacher models.

657

658

A.2 HOW TO DETERMINE  $\alpha$  AND  $\beta$ 659  
660A.2.1 HOW TO DETERMINE  $\alpha$ 661  
662  
663  
664  
665  
666  
667  
668

For  $\alpha$ , it is used to filter out relatively small probability differences that would otherwise affect the analysis. To obtain a relatively objective threshold, we asked 10 annotators to judge the magnitude of probability differences. For each annotation, the annotators were shown an output-probability line chart similar to that in Figure 2. Given all the sentences in the response, we asked them to identify those whose probability differences could be regarded as relatively negligible, such as sentences 55–65 in Figure 2. After we explained the definitions of the four sentence types and the goal of our analysis, the annotators manually selected the sentences that appeared more likely to be Shared Sentences, and we then derived the final rounded value of  $\alpha = 0.1$  from their judgments.

Figure 8: Illustration of  $\beta$  selection for Deepseek-Distill-Qwen-7B on AIME24.

679

680

A.2.2 HOW TO DETERMINE  $\beta$ 681  
682  
683

The parameter  $\beta$  serves primarily to facilitate a clearer separation of the relative proportions of different action types, enabling more effective provenance analysis. Specifically, for each trajectory, we compute the proportion of sentence attributed to each action type (e.g., in trajectory 1, 60% of sentences are labeled as Teacher Sentence; in trajectory 2, 70%). We then construct histograms showing the distribution of these proportions across all samples for each action type (e.g., plotting the distribution of Teacher Sentence proportions using values such as [0.6, 0.7, ...]).

Figure 9: Illustration of  $\beta$  selection process.

Intuitively, we choose  $\beta$  to maximize the separation between the distributions of different action types, so that their overlap is minimized, particularly for the action

694  
695  
696  
697  
698  
699  
700  
701

types that dominate in frequency. As shown in Figure 8, for instance, when  $\beta = 0.15$ , the histogram reveals that most samples exhibit a clear predominance of Teacher Sentence, providing a clean characterization of the model’s output behavior. In contrast, when  $\beta = 0.2$ , the distributions of Teacher Sentence and Boosted Sentence exhibit significant overlap, indicating that samples within the overlapping region are sensitive to minor fluctuations and may be ambiguously classified. A similar problem also exists with  $\beta = 0.1$ .

702 **Algorithm 1** Search for Optimal  $\beta$ 


---

703  
**Require:** teacher model  $M_T$ , student model  $M_S$ , batch of data  $\mathcal{D}$   
**Ensure:** optimal threshold  $\beta^*$

704 1: Feed  $\mathcal{D}$  into  $M_T$  and  $M_S$  to obtain sentence-level output probabilities.  
705 2: Following Section 4.2.2, partition sentences and get sentence-level probabilities into: Common  
706 Sentence set  $\mathcal{C}$ , Teacher Sentence set  $\mathcal{T}$ , Student Sentence set  $\mathcal{S}$ .  
707 3: best overlap:  $O^* \leftarrow +\infty$   
708 4: best  $\beta$ :  $\beta^* \leftarrow \text{None}$   
709 5: **for**  $\beta \in \{0.05, 0.10, \dots, 1.0\}$  **do**  
710 6:   Compute the histogram overlap between  $\mathcal{S}$  and  $\mathcal{C}$  under threshold  $\beta$ , denote it as  $O_1$ .  
711 7:   Compute the histogram overlap between  $\mathcal{C}$  and  $\mathcal{T}$  under threshold  $\beta$ , denote it as  $O_2$ .  
712 8:   Current overlap  $O \leftarrow O_1 + O_2$   
713 9:   **if**  $O < O^*$  **then**  
714 10:      $O^* \leftarrow O$   
715 11:      $\beta^* \leftarrow \beta$   
716 12:   **end if**  
717 13:   **if**  $\text{mean}(\mathcal{C}) > \text{mean}(\mathcal{T})$  **then**  
718 14:     **break**  
719 15:   **end if**  
720 16: **end for**  
721 17: **return**  $\beta^*$

---

722  
723  
724  
725  
726 Therefore, for Deepseek-Distill-Qwen-7B on AIME24, we set  $\beta = 0.15$ . For the experiments in  
727 Section 3.3, we evaluate all values of  $\beta$  in the range  $[0.05, 0.2]$  with a step size of 0.05. The  
728 final selected thresholds are:  $\beta = 0.1$  for Deepseek-Distill-Qwen-7B on GPQA-D,  $\beta = 0.1$   
729 for DeepSeek-R1-0528-Qwen3-8B on AIME24,  $\beta = 0.15$  for DeepSeek-R1-0528-Qwen3-8B on  
730 GPQA-D,  $\beta = 0.1$  for LIMO-v2 model on AIME24 when using QwQ-32B as teacher model,  
731  $\beta = 0.2$  for LIMO-v2 model on GPQA-D when using QwQ-32B as teacher model,  $\beta = 0.15$  for  
732 LIMO-v2 model on AIME24 when using Deepseek-R1 as teacher model, and  $\beta = 0.1$  for LIMO-v2  
733 model on AIME24 when using Deepseek-R1 as teacher model.

734 In fact, threshold selection is also an adaptive procedure rather than a manually specified parameter  
735 that must be tuned via repeated training runs, which makes it fundamentally different from hyperpa-  
736 rameter selection in neural networks. Taking the training-time pipeline as an example, we perform  
737 the search using Algorithm 1. For the four training configurations in Table 1, the chosen  $\beta$  values  
738 are 0.1, 0.1, 0.15, and 0.2, respectively. In addition, we show how this algorithm selects the value  
739 0.2 for Table 1, Setting 4, as illustrated in Figure 9.

740  
741  
742  
743 A.2.3 SENSITIVITY ANALYSIS  
744  
745

746 We also illustrate threshold selection on first training setting (DeepSeek-R1 + Qwen3-4B-Base +  
747 AceReason-1.1-SFT) in Figure 6. The corresponding quantitative results are reported in Table 3.  
748 As can be seen, the threshold identified as optimal before training also yields the best performance  
749 after training. The post-training metrics in Table 3 and the pre-training visualizations in Figure 6  
750 demonstrate that the proposed method is purely data-driven, requires no additional training, and is  
751 practically applicable. Near-optimal  $\beta$  also outperform the Vanilla baseline, indicating that perfor-  
752 mance is not highly sensitive to the choice of  $\beta$ .

753 Additionally, we examine the sensitivity of our analysis to the choice of  $\beta$ . As shown in Figure 8,  
754 the relative trends and distinctions in action-type output probability between correct and incorrect  
755 samples remain consistent across different  $\beta$  values, suggesting that our main conclusions are robust  
to the  $\beta$  setting.

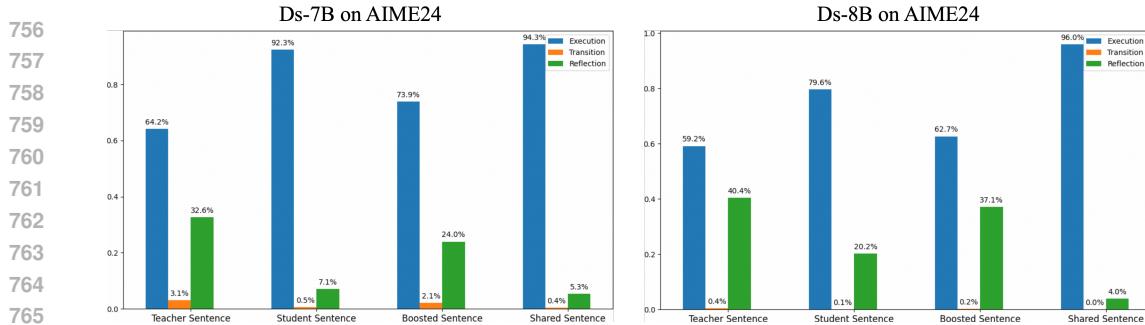


Figure 10: Differences across action types. Using SEAL’s behavioral categorization, we classify all actions and report the proportion of behavior types within each action type.

### A.3 MORE ANALYSIS RESULTS

#### A.3.1 ALTHOUGH DISTILLED MODELS GENERALLY EXHIBIT SUBSTANTIALLY STRONGER LONG-CONTEXT GENERATION CAPABILITIES, THE SOURCES OF THESE IMPROVEMENTS ARE NOT UNIFORM ACROSS MODELS.

As shown in Figure 3, within DeepSeek-Distill-Qwen-7B, once the sequence length exceeds approximately 4K tokens, the generated outputs are quickly transformed into Teacher Sentence. In contrast, this behavior is not observed in DeepSeek-R1-0528-Qwen3-8B or LIMO-v2 model. We attribute this discrepancy to differences in the underlying student model: DeepSeek-Distill-Qwen-7B is distilled from Qwen2.5-Math-7B (Yang et al., 2024), which has an effective context length of approximately 4K tokens. Beyond this limit, the model’s outputs predominantly reflect patterns inherited from the teacher model. By comparison, DeepSeek-R1-0528-Qwen3-8B and LIMO-v2 model are based on Qwen3-8B-Base (Team, 2025a) and Qwen2.5-32B-Instruct, respectively. These models support substantially longer effective context lengths and thus avoid this limitation.

#### A.3.2 WHAT BEHAVIORS ARE INCLUDED IN DIFFERENT ACTION TYPES?

We categorize each action using SEAL (Chen et al., 2025), which defines three types of behaviors: execution, reflection, and transition. Execution refers to steps that directly advance problem solving, reflection denotes verification, checking, or questioning of the existing reasoning process, and transition represents an intentional change in the current reasoning direction or strategy. Special tokens are excluded from the statistics. The results are shown in Figure 10, from which we make the following three observations.

(1) Both student models exhibit a certain degree of reflection and transition, and reasoning distillation further activates and strengthens these behaviors. Specifically, Student Sentences already contain reflection and transition behaviors, while Boosted Sentences display more of these behaviors than Student Sentences and Shared Sentences after distillation. Although there remains a gap between RL and reasoning distillation, this observation is similar with prior work (Wang et al., 2025; Gandhi et al., 2025), which suggests that reflection and transition abilities are already latent in student models and that RL/Reasoning distillation as a post-training method serves to activate them. (2) The vast majority of Shared Sentences are Execution. (3) The student model of Deepseek-R1-0528-Qwen3-8B (Ds-8B) shows stronger reflective ability than the student model of Deepseek-Distill-Qwen-7B (Ds-7B), as evidenced by the higher proportion of reflection behaviors in Student Sentences.

#### A.3.3 TEACHER SENTENCES ACROSS DOMAINS

As illustrated in Figure 11, we further examine the characteristics of Teacher Sentences across domains. In mathematics, Teacher Sentences primarily consist of explicit mathematical operations, followed by checks of these operations and validations of the solution logic. In science, they more often involve inspections of reasoning chains and targeted recall of relevant knowledge. Despite these differences, we observe common patterns: when reasoning stalls, Teacher Sentences prompt the recall of key facts and encourage reflective adjustments to the reasoning process. We hypothe-

<pre> 810 &lt;think&gt; 811 In first, I need to synthesize 5-isopropyl-3,4-dimethylcyclohex-1-ene using ring-closing metathesis (RCM). 812 RCM is a reaction that forms a ring from a diene by eliminating ethene, typically requiring a diene with specific stereochemistry or substitution to form the desired ring size. 813 The target molecule is 5-isopropyl-3,4-dimethylcyclohex-1-ene. 814 It's a cyclohexene ring with a double bond at position 1, and substituents at positions 3,4, and 5: isopropyl at 5, and methyl groups at 3 and 4. 815 In ring-closing metathesis, the diene precursor should be such that when it undergoes metathesis, it forms a six-membered ring with the correct substituents and the double bond in the correct position. 816 The target is a cyclohexene, so the ring-closing should form a six-membered ring. 817 In RCM, the ring size is determined by the number of atoms in the chain between the two ends of the diene. 818 For a six-membered ring, the diene should have a chain with four atoms or groups that will form the ring, but in metathesis, it's the diene that closes the ring. 819 Recall how RCM works: for example, a diene like 1,6-heptadiene can form cyclohexene. 820 The general idea is that the diene has two double bonds, and metathesis between them forms a ring and ethene. 821 For a six-membered ring cyclohexene, the diene should be such that the two ends are connected by a chain that, when the double bonds react, forms the six-member ring. 822 In standard RCM, for a cyclohexene, the diene is typically a 1,6-diene with substituents on the chain. 823 But in this case, the target is 5-isopropyl-3,4-dimethylcyclohex-1-ene. 824 Let's number the ring. 825 Assume the cyclohexene ring has the double bond between C1 and C2. 826 Then substituents: methyl at C3 and C4, isopropyl at C5. 827 In standard naming, cyclohexene with double bond between C1-C2, so C1 and C2 are sp2, C3-C6 are sp3. 828 Substituents: methyl at C3 and C4, isopropyl at C5. 829 Now, for RCM to form this ring, the diene precursor should have the two ends that will become C1 and C2 of the ring, and the chain connecting them should be such that after metathesis, the ring is formed. 830 In RCM, the double bonds are internal, and metathesis forms a new double bond in the ring. 831 Typically, for a cyclohexene via RCM, the diene is a 1,6-diene where the two ends are the groups that will be part of the ring. 832 For example, to make cyclohexene, you might have a diene like <chem>CH2=CH-CH2-CH2-CH=CH2</chem>, but that's symmetric. 833 But here, there are substituents. 834 So, for the ring-closing to form a six-member ring, the diene should have the two double bonds separated by three atoms. 835 In a six-member ring, there are six atoms. 836 When you form the ring via RCM, the two ends of the diene become adjacent in the ring. 837 Standard way: for RCM to give a cyclohexene, the diene should be a 1,6-diene with the two ends being the groups that are not part of the double bonds. </pre>	<b>Ds-8B on GPQA-D</b>
<pre> 838 &lt;think&gt; 839 In the problem involves a triangle with circumcenter O and incenter I. 840 IA is perpendicular to OI, circumradius is 13, and inradius is 6. 841 I need to find the product AB times AC. 842 First, I recall some properties of the circumcenter and incenter. 843 The circumcenter O is the center of the circumcircle, which passes through all three vertices, and its radius is given as 13. 844 The incenter I is the center of the incircle, tangent to all three sides, with radius 6. 845 Also, IA is perpendicular to OI. 846 IA is the line from I to A, and it's perpendicular to the line joining O and I. 847 I need to find AB * AC, which is the product of two sides from A. 848 I remember that in any triangle, the distance between O and I is given by <math>\sqrt{OI^2 = R(R - 2r)}</math>, where R is the circumradius and r is the inradius. 849 Let me verify that. 850 Yes, the standard formula is <math>\sqrt{OI^2 = R(R - 2r)}</math>. 851 Here R=13, r=6, so <math>\sqrt{OI^2 = 13(13 - 12)} = 13</math> (times 1 = 13). 852 So OI is <math>\sqrt{13}</math>. 853 Now, there's a condition that IA is perpendicular to OI. 854 IA is the distance from I to A, but it's a line, so probably the line IA is perpendicular to the line OI. 855 In the context, it says "IA perpendicular to OI", so I think it means the line segments or lines are perpendicular. 856 Now, I need to relate that to the sides. 857 I recall that in a triangle, the distance from incenter to a vertex can be expressed in terms of the sides and angles. 858 Let me denote the triangle as ABC, with sides opposite to A, B, C as a, b, c respectively. 859 So AB = c, AC = b, BC = a. </pre>	<b>Ds-8B on AIME24</b>

Figure 11: Comparison of Teacher Sentences across domains. Teacher Sentences are highlighted in light green.

Table 5: Training config.

	Qwen3-4B-Base	Qwen3-8B-Base	Qwen3-4B-Instruct-2507	Qwen2.5-7B-Instruct
learning_rate	5e-5	5e-5	5e-5	8.0e-5
cutoff_len	32k	64k	64k	16k
epoch	6	6	6	5
batchsize	32	32	32	512
lr_scheduler_type	cosine_with_min_lr	cosine_with_min_lr	cosine_with_min_lr	cosine
min_lr	1e-5	1e-5	1e-5	0
warmup_ratio	0.1	0.1	0.1	0.1

size that these shared behaviors explain why training on scientific-domain data containing a higher proportion of Teacher Sentences can also benefit performance on mathematics test sets.

#### A.4 MORE DISCUSSIONS

##### A.4.1 A NEW PERSPECTIVE ON UNDERSTANDING MODELS

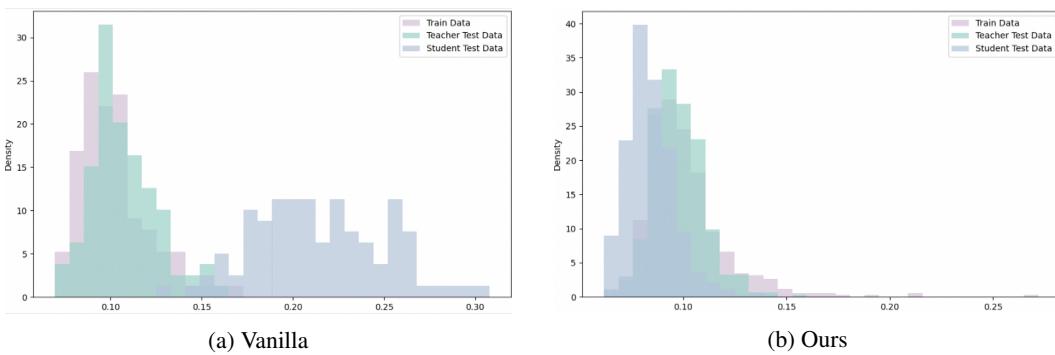
Most existing approaches Lee et al. (2025); Liu et al. (2025a) to understanding model outputs operate in a human-centric semantic space. They typically seek to decompose intermediate activations or outputs into human-interpretable units, such as natural language concepts, symbolic structures, or predefined semantic categories. While this line of work has yielded valuable insights, it implicitly assumes that model behavior is best understood by mapping it onto human-designed semantic structures.

We agree that semantic/structure-level analyses are important. But we believe that decomposing the outputs of a reasoning model into human-interpretable semantic units is not the only way to

understand model behavior. Instead, our method offers a novel and complementary perspective: we construct a decomposition that is naturally induced by the model itself and directly ties its outputs to performance on the test set. This model-centric view allows us to analyze how different components of the model’s output contribute to its empirical performance, without requiring a predefined set of human-interpretable semantic units.

#### 870 A.4.2 ALIGNMENT OF OUTPUT DISTRIBUTIONS

872 **In the distillation setting, we define better reasoning as closer alignment with the teacher**  
 873 **model’s output distribution. The goal of reasoning distillation is therefore twofold: (i) the**  
 874 **student model should learn the teacher model’s output distribution, and (ii) this alignment**  
 875 **should yield better performance on the test set.** In addition to the existing analysis of the second  
 876 point (in main text), we further examine the first point. To examine (i), we construct three sets of  
 877 sentences: the teacher model’s outputs on the training set, the teacher model’s outputs on the test  
 878 set, and the distilled student model’s outputs on the test set. For each set, we feed every response  
 879 into both the teacher model and the distilled student model, compute the difference between the  
 880 output probabilities for each sentence, and then aggregate these differences to obtain the empirical  
 881 distribution histogram of probability discrepancies.



893 **Figure 12: Distribution of output probability differences.** “Train Data”, “Teacher Test Data”, and  
 894 “Student Test Data” represent the outputs of the teacher model on the training set, the outputs of  
 895 the teacher model on the test set, and the outputs of the distilled student model on the test set,  
 896 respectively.

897 The results, shown in the Figure 12, compare two different data-filtering strategies. Under both  
 898 strategies, the probability discrepancy is small on the training data (the teacher’s outputs on the  
 899 training set), which is expected because the training data are well fitted. The discrepancy on the  
 900 teacher’s outputs on the test set is slightly larger, indicating that, although the queries in the training  
 901 and test sets differ, the student model still largely follows the teacher model’s output distribution  
 902 when evaluated on the teacher’s context. Remarkably, when evaluated on the distilled model’s own  
 903 context on the test set, our method achieves an even smaller probability discrepancy than on the training  
 904 data. This strongly suggests a high degree of alignment with the teacher’s output distribution.  
 905 It is important to note that the teacher distribution is not directly observable. We only approximate  
 906 it indirectly via the probabilities assigned to sentences under different contexts. The fact, that the  
 907 smallest probability discrepancy occurs on the distilled model’s own outputs on the test set, indicates  
 908 that our method allows the student model, in its own generation context, to closely align with the  
 909 teacher model’s output distribution.

#### 911 A.5 TRAINING DETAILS

913 For the OpenThought3-1.2M dataset, we used only the mathematical problems and randomly sam-  
 914 pled 50k questions. For each question, the official dataset provides 16 responses generated by QwQ;  
 915 from these, we randomly selected 8 responses to form our initial training set.

917 For the AceReason-1.1-SFT dataset, we first randomly sampled 50k mathematical questions. For  
 each question, the official dataset provides multiple responses generated by R1. We then addition-

918 ally generated 1–5 responses per question using GPT-OSS-120B (configured with Reasoning: 919 high). Together, the R1 and GPT-OSS-120B responses served as our initial training set. 920

921 The training configuration is provided in Table 5.

922 **A.6 RELATED WORK**

923 **A.6.1 LLM DISTILLATION**

926 **Knowledge Distillation.** Knowledge distillation was first introduced by (Hinton et al., 2015a) as a 927 technique for transferring the dark knowledge of a teacher model to a lightweight student model via 928 soft labels, thereby enabling substantial model compression while preserving most of the original 929 accuracy. Subsequent works (SE, 2014; Zagoruyko & Komodakis, 2016) has extended this framework 930 from multiple perspectives, for example by distilling feature layers or intermediate representations 931 and by introducing relational or structural distillation, so that the student not only mimics the output 932 distribution but also aligns the structural properties of the hidden representation space. Later works 933 (Ko et al., 2024; 2025) further improve alignment by approximating sequence-level KL divergence, 934 leading to more effective distillation. In practice, knowledge distillation (Gou et al., 2021) has been 935 widely applied to tasks such as image classification, object detection, and natural language processing, 936 and has become one of the mainstream approaches for model compression and acceleration.

937 **Reasoning Distillation.** Distilling the reasoning abilities of large language models has been an im- 938 portant problem since their emergence (Kim & Rush, 2016; Ho et al., 2023; Ko et al., 2024; 2025). 939 Before the advent of large-scale reasoning models such as O1 (Jaech et al., 2024), traditional 940 reasoning distillation methods primarily transferred capabilities by aligning intermediate features or 941 output probabilities between teacher and student models (Kim & Rush, 2016; Ho et al., 2023; Ko 942 et al., 2024; 2025). To teach reasoning more explicitly, prior work (Hsieh et al., 2023; Kim & Rush, 943 2016; Ho et al., 2023; Ko et al., 2024; 2025) constructs responses that include detailed reasoning 944 traces and trains the student on these signals, thereby strengthening its mastery of reasoning. In 945 the era of large-scale reasoning models such as R1 (DeepSeek-AI, 2025) and QwQ (Team, 2025b), 946 which naturally exhibit chain-of-thought reasoning and achieve strong performance, distilling their 947 capabilities into smaller models has become an effective and practical path toward improved effi- 948 ciency. DeepSeek (DeepSeek-AI, 2025) pioneered this line of work by showing that supervised 949 fine-tuning on the outputs of a reasoning teacher, which is also the approach we focus on in this 950 paper, can dramatically enhance the reasoning abilities of smaller models. Numerous subsequent 951 projects (e.g., OpenR1 (Hugging Face, 2025), OpenThoughts (Guha et al., 2025), a-m-team (Zhao 952 et al., 2025), NVIDIA AceReason (Liu et al., 2025b), OmniThought (Guha et al., 2025), LIMO (Ye 953 et al., 2025), DeepMath (He et al., 2025)) have devoted substantial effort to constructing and refining 954 large-scale corpora of challenging reasoning problems paired with teacher responses, using rigorous 955 quality filtering, correctness checks, and diversity-aware curation. Most recently, GRAPE (Zhang 956 et al., 2025) scores candidate responses with the student model and preferentially selects examples 957 whose likelihoods best match the student’s current distribution, thereby steering training toward data 958 that is already well aligned with the student. **Rather than focusing solely on artificially designed 959 rules and heuristic rules, we view reasoning distillation as a capability-transfer problem from 960 teacher to student. We aim to quantify the sources of a distilled model’s capabilities: given a 961 context, which actions in a trajectory from distilled model are more likely to originate from 962 the teacher’s behavior rather than the student’s existing tendencies? Building on this perspec- 963 tive, we introduce a data selection criterion that jointly compares teacher–student output 964 distributions and focuses on sentences whose probabilities indicate stronger teacher-originated 965 behavior. This provenance-aware criterion complements prior student-only selection in the 966 following way: it provides an explicit cross-model signal for reasoning transfer.** In Section 4, 967 we show that provenance-aware selection outperforms student-only alignment in our settings.

968 **A.6.2 MODEL AUDITING**

969 Another closely related area is model auditing, a growing line of work that studies (Song & 970 Shmatikov, 2019; Carlini et al., 2022; Dekoninck et al., 2024) auditing generative models to un- 971 derstand what data they memorize and to attribute outputs back to underlying data sources. For 972 example, prior work (Song & Shmatikov, 2019) shows that rare tokens in the training data tend to 973 be memorized by text generation models, and uses shadow models together with an audit classifier

972 (e.g., an SVM on token-rank features) to distinguish whether a user’s data was included in training.  
973 Separately, subsequent work (Carlini et al., 2022) formalizes extractability as the ability of a model,  
974 given a prefix, to greedily regenerate the exact suffix from the training set, and systematically studies  
975 how repetition and sequence length affect the fraction of such extractable sequences. **In contrast,**  
976 **our work targets model-level provenance in a distillation setting: rather than asking whether**  
977 **specific data are memorized, we aim to trace which upstream models are the sources of a given**  
978 **output, shifting the focus from data membership to the lineage of the models themselves.**

979

## 980 A.7 LLM USAGE

981

982 We used Qwen3 for polishing, followed by manual refinement.

983

984

985

986

987

988

989

990

991

992

993

994

995

996

997

998

999

1000

1001

1002

1003

1004

1005

1006

1007

1008

1009

1010

1011

1012

1013

1014

1015

1016

1017

1018

1019

1020

1021

1022

1023

1024

1025