Which Reasoning Trajectories Teach Students to Reason Better? A Simple Metric of Informative Alignment

Yuming Yang[1,2], Mingyoung Lai[3], Wanxu Zhao[1], Xiaoran Fan[1], Zhiheng Xi[1], Mingqi Wu[1], Chiyue Huang[4], Jun Zhao[1], Haijun Lv[2], Jian Tong[2], Yunhua Zhou[2], Yicheng Zou[2]*, Qipeng Guo[2], Tao Gui[5,6], Qi Zhang[1,5,6]*, Xuanjing Huang[1,5,6]

[1]College of Computer Science and Artificial Intelligence, Fudan University
[2]Shanghai AI Laboratory [3]University of Toronto [4]University of Sydney
[5]Institute of Trustworthy Embodied AI, Fudan University
[6]Shanghai Key Laboratory of Multimodal Embodied AI

yumingyang23@m.fudan.edu.cn, zouyicheng@pjlab.org.cn, qz@fudan.edu.cn

Abstract

Long chain-of-thought (CoT) trajectories provide rich supervision signals for distilling reasoning from teacher to student LLMs. However, both prior work and our experiments show that trajectories from stronger teachers do not necessarily yield better students, highlighting the importance of data-student suitability in distillation. Existing methods assess suitability primarily through student likelihood, favoring trajectories that align closely with the student model's current behavior but overlooking more informative ones. Addressing this, we propose *Rank-Surprisal Ratio* (RSR), a simple metric that captures both alignment and informativeness to assess the suitability of a reasoning trajectory. RSR is motivated by the observation that effective trajectories typically balance learning signal strength and behavioral alignment by combining low absolute probability with relatively high-ranked tokens under the student model. Concretely, RSR is defined as the ratio of a trajectory's average token-wise rank to its average negative log-likelihood, and is straightforward to compute and interpret. Across five student models and reasoning trajectories from 11 diverse teachers, RSR strongly correlates with post-training reasoning performance (average Spearman 0.86), consistently outperforming existing metrics. We further demonstrate its practical utility in both trajectory selection and teacher selection. Code and data are available at https://github.com/UmeanNever/RankSurprisalRatio.

1 Introduction

Recent advances in reasoning-oriented large language models (LLMs) are largely driven by their ability to generate long chain-of-thought (CoT) trajectories (Wei et al., 2022; Zhang et al., 2025c). Beyond enabling complex inference at test time, such trajectories also provide powerful supervision signals for training student models (Guo et al., 2025; Muennighoff et al., 2025) or cold-starting reinforcement learning (Yang et al., 2025a) through supervised fine-tuning (SFT).

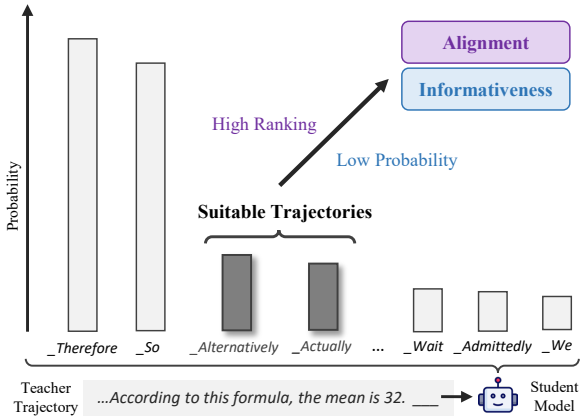

Figure 1: Illustration of the intuition behind *Rank-Surprisal Ratio*. Suitable reasoning trajectories should balance informativeness and alignment, having low absolute probability while their tokens remain relatively high-ranked under the student model.

Yet, stronger reasoning teachers do not necessarily yield better students (Li et al., 2025b; Guha et al., 2025). Our extensive experiments show that the post-training effectiveness of reasoning trajectories varies substantially across student models, indicating that the suitability between data and student is critical for effective learning. Existing data engineering methods assess data suitability primarily through the student's probability assignments (Zhang et al., 2025b; Just et al., 2025), favoring high-likelihood trajectories that align closely with the model's current behavior. Such trajectories, however, often provide limited new learning signals. In contrast, more informative trajectories are typically less familiar to the student and thus overlooked by these methods. To facilitate more effec-

*Corresponding Authors.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 42123–42150
July 2-7, 2026

tive learning, it is crucial to strike a balance between familiarity and informativeness, echoing the psychological concept of the zone of proximal development (Vygotsky and Cole, 1978). This leads to a fundamental *Informative Alignment* challenge: **how to identify reasoning data that are both well aligned with the student and sufficiently informative?**

To address this challenge, we propose a simple yet effective metric, *Rank-Surprisal Ratio* (RSR), which quantifies the suitability of a reasoning trajectory for a given student by jointly capturing alignment and informativeness. Motivated by our preliminary analysis, we argue that the dilemma between providing new signals and aligning with student's existing behavior can be resolved by trajectories exhibiting both absolute unfamiliarity and relative familiarity. Concretely, effective trajectories should deviate from the student's own generations, receiving **low absolute probability** under the student model, while remaining compatible with its overall generation patterns, such that their tokens still **rank relatively high** in the model's prediction distribution over the vocabulary (Figure 1).

Based on this insight and consistent numerical patterns observed in simulation studies, we define our suitability metric, *Rank-Surprisal Ratio*, as the ratio between a trajectory's average token-wise rank[1] and its average negative log-likelihood (surprisal). RSR can be computed with a single forward pass, requires no additional verifier or test data, and is straightforward to interpret. Lower RSR indicates better informative alignment, identifying trajectories that are both informative and well aligned with the student.

We validate the effectiveness of *Rank-Surprisal Ratio* through correlation analyses on 5 student LLMs using math reasoning trajectories generated by 11 representative teacher models. Across all students, the RSR of trajectories exhibits a strong correlation with post-training performance, achieving an average Spearman correlation of 0.86 and consistently outperforming alternative metrics. Furthermore, to explore its practical value in data engineering, we apply RSR to trajectory selection and teacher selection. Our experiments show that RSR not only selects more effective training trajectories for each problem from candidates generated by diverse teachers, but also identifies more suitable teacher models using only a small amount of data, consistently outperforming existing selection methods across all five students in both settings.

Our main contributions are three-fold:

- We present a systematic distillation study across a wide range of teacher and student models, showing that the effectiveness of reasoning trajectories differs across students and highlighting the importance of data-student suitability (§ 2).
- We propose *Rank-Surprisal Ratio*, a simple metric that quantifies the suitability of a reasoning trajectory for a given student model by jointly capturing alignment and informativeness (§ 3), achieving a strong correlation with post-training performance (§ 4).
- We demonstrate the practical utility of RSR in two data engineering scenarios, trajectory selection and teacher selection, where it serves as an effective criterion and outperforms existing methods (§ 5). We further discuss the broader applicability of RSR to non-CoT data and subset selection (§ 7).

2 The Need for Student-Specific Data

To understand which types of reasoning trajectories most effectively improve student models after SFT, we conduct a comprehensive large-scale study involving five widely adopted student models and eleven diverse reasoning-oriented teacher models, yielding 55 teacher-student pairings. We perform SFT experiments for each of these pairs.

2.1 Experimental Settings

Our teacher-student pairing study involves two major steps: (1) For each teacher model, we prompt it to generate a long CoT response for each math problem in our 5000-problem set (see § A.1), forming a trajectory dataset specific to that teacher. (2) For each teacher-student pair, we fine-tune the student model on the corresponding teacher dataset and evaluate its reasoning performance. All students are pre-trained base models. To reduce variance induced by stochastic trajectory sampling, we perform three independent generation runs for each teacher and conduct SFT separately on each resulting dataset for every teacher-student pair. Reported results are averaged over these three runs. More implementation details are provided in Appendix A.

Teachers We use 11 reasoning LLMs (§ A.3) spanning 4B to 671B parameters across multiple model families, including DeepSeek (Guo et al.,

[1]Higher-ranked tokens have lower rank values.

Teacher Models	Params	Student Models (Base)					Teacher Performance
		Qwen-3-14B	LLaMA-3.1-8B	Qwen-2.5-7B	Qwen-3-4B	Qwen-2.5-3B	
Deepseek-R1	671B	77.1	28.1	47.3	55.8	29.6	91.1
Qwen-3-235B-Thinking	235B	71.8	22.0	45.0	53.4	26.4	91.2
GPT-OSS-120B	120B	66.7	15.2	40.7	47.9	22.9	88.3
Nemotron-Super	49B	72.2	23.7	48.3	56.4	33.0	82.3
QwQ-32B	32B	77.4	27.1	52.0	61.2	33.0	85.2
Qwen-3-30B-Thinking	30B	77.2	26.7	50.0	58.8	31.2	92.3
Magistral-Small	24B	68.8	22.8	47.6	52.2	30.6	71.0
GPT-OSS-20B	20B	69.5	17.9	42.7	48.4	24.4	83.4
Phi-4-Reasoning-Plus	14B	54.1	14.5	35.2	40.2	18.2	72.7
Qwen-3-8B	8B	74.6	26.5	52.0	61.2	34.2	82.5
Qwen-3-4B-Thinking	4B	76.8	28.2	51.8	61.9	33.3	87.3

Table 1: Distillation results showing post-training reasoning performance of student models trained on trajectories from different teacher models, evaluated by average Acc@4 on AIME'25, AIME'24, AMC'23, and MATH500. Darker and lighter shading indicate the best and second-best results, respectively. Student performance varies significantly across teacher-student pairs, highlighting the importance of data-student suitability.

2025), GPT-OSS (Agarwal et al., 2025), Qwen (Yang et al., 2025a), LLaMA-Nemotron (Bercovich et al., 2025), and Phi (Abdin et al., 2025).

Benchmarks We evaluate the reasoning performance of fine-tuned student models on four popular math benchmarks (AIME'25, AIME'24, AMC'23, and MATH500 (Hendrycks et al., 2021)) using the Acc@4 metric (§ A.4), and report results averaged across all benchmarks.

2.2 Results

Table 1 presents the results of our teacher-student pairing distillation study, revealing that:

Stronger teachers do not necessarily produce better students. Teacher capability, whether measured by parameter scale or reasoning performance, does not reliably predict student improvement. For example, the 671B and 235B models often underperform smaller teachers such as QwQ-32B on multiple students. Similarly, teachers with strong reasoning performance do not consistently yield the best outcomes for all student models.

Data-student suitability is critical for eliciting reasoning improvements. The effectiveness of teacher trajectories is highly student-specific and depends critically on their suitability for the student model. Pairing strong teachers (Deepseek-R1) with much weaker students (Qwen-2.5-3B) often fails to yield strong performance, while weaker teachers (Qwen-3-4B-Thinking) can likewise be ineffective when training stronger students (Qwen-3-14B). Moreover, teachers from distant model families (GPT-OSS) often lead to inferior results, suggesting that unfamiliar reasoning patterns are harder for students to absorb. Overall, we find no simple teacher-student pairing rule based on surface attributes such as parameter scale or model family, indicating that reasoning data suitability is a nuanced property requiring deeper investigation.

3 Measuring Data-Student Suitability

In this section, we explore metrics for measuring data-student suitability, with the goal of jointly capturing informativeness and alignment. We begin by introducing two fundamental token-level measures: surprisal and rank (§ 3.1), and analyzing the limitations of existing probability-based metrics (§ 3.2). We then abstract our insights on suitable reasoning data and conduct simulation studies to identify quantitative patterns (§ 3.3). Finally, we propose our trajectory-level metric (§ 3.4).

3.1 Surprisal and Rank

We first introduce two fundamental methods for quantifying the amount of information a token carries with respect to a student model. They serve as the building blocks of our metric.

Surprisal (Negative Log-Likelihood) A common measure is based on the probability of generating the current token t_k given its preceding context $\mathbf{c}_k = (t_1, \ldots, t_{k-1})$ under the student model θ. For numerical stability, probabilities are typically transformed into log space, and the negative log-likelihood—also known as *surprisal*—is used as a measure of informativeness (Hale, 2001).

$$\text{Surprisal}(t_k) = -\log p_\theta(t_k \mid \mathbf{c}_k) \quad (1)$$

Rank Another method considers the rank of the current token within the model's prediction distribution over the vocabulary $\mathcal{V}$. Formally, given the con-

Teacher Models	Student Performance ↑	Probability-based Metrics ↓		Rank-Surprisal Metrics ↓		
		Avg-Surprisal	Avg-$\text{Surp}_{\text{local}}$	Avg-$\text{RSR}_{\text{token}}$	Avg-$\text{RSR}^{\text{filter}}_{\text{token}}$	**RSR (Ours)**
Qwen-3-8B	52.0	0.65	1.16	$2.01_{\times 10^7}$	3.15	2.89
Qwen-3-30B-Thinking	50.0	0.77	1.27	$2.25_{\times 10^7}$	3.47	2.95
Nemotron-Super	48.3	0.60	1.04	$5.51_{\times 10^7}$	3.95	3.08
Deepseek-R1	47.8	0.83	1.35	$2.98_{\times 10^7}$	3.31	3.00
Magistral-Small	47.6	0.55	1.03	$3.58_{\times 10^7}$	4.38	3.09
GPT-OSS-20B	42.7	1.36	1.78	$5.15_{\times 10^6}$	11.10	3.83

Table 2: Comparison of the student's post-training performance and data-student suitability metrics across trajectories from different teacher models, evaluated on Qwen-2.5-7B. Darker shading indicates higher performance or better suitability. Metrics whose trends align with performance (e.g., RSR) provide more reliable suitability estimates. Complete metric scores for all teacher and student models are provided in § F.

ditional distribution $p_\theta(\cdot \mid \mathbf{c}_k)$, the rank of token t_k is defined as the number of tokens with strictly higher probability (Ravichander et al., 2025).

$$\text{Rank}(t_k) = 1 + \sum_{t' \in \mathcal{V}} \mathbb{I}\big[p_\theta(t' \mid \mathbf{c}_k) > p_\theta(t_k \mid \mathbf{c}_k)\big] \tag{2}$$

Unlike surprisal, rank captures relative familiarity of the token by comparing target token against alternative candidates, revealing signals overlooked by probability-based measures. For instance, a token may be assigned a low absolute probability while still ranking among the top candidates.

3.2 Limitations of Probability-Based Metrics

Existing work primarily relies on log-probability or surprisal to assess data suitability. For example, Zhang et al. (2025b) select trajectories based on the average log-probability of response tokens under the student model. Since surprisal is the negation of log-probability, we implement this metric as the average surprisal, denoted as *Avg-Surprisal*. More recently, Just et al. (2025) compute token-level log-probability based on a local context $\mathbf{c}_k^{\text{local}}$ (several preceding sentences), which we implement as average local surprisal (*Avg-*Surp_{local}, § B.6). Under these metrics, trajectories with lower surprisal are considered more suitable for the student model.

However, as shown in Table 2, lower surprisal (i.e., higher probability) does not necessarily lead to better post-training reasoning performance. Both Avg-Surprisal and Avg-$\text{Surp}_{\text{local}}$ assign lower surprisal to trajectories from Nemotron-Super and Magistral-Small, yet training on such trajectories fails to improve reasoning performance. Similar patterns are observed across all student models, indicating that probability-based metrics tend to favor data that are familiar but insufficiently informative.

At the other extreme, trajectories with very high surprisal (e.g., GPT-OSS-20B) also perform poorly. In contrast, trajectories with moderate surprisal values (e.g., Qwen-3-8B) achieve better results. This observation motivates us to investigate the mechanism underlying the surprisal trade-off.

3.3 Insight and Simulation

The above analysis suggests that suitable (i.e., effective) teacher trajectories should strike a balance between data informativeness and alignment with the student's current behavior: they should be neither overly similar to the student's own generations nor excessively deviant from its prediction distribution.

At first glance, this balance may appear to pose a dilemma. However, we argue that it can be resolved by viewing informativeness and alignment through the lens of *absolute unfamiliarity and relative familiarity*. Informativeness does not require trajectories to be entirely unfamiliar; rather, it suffices that they deviate from the dominant patterns and thus have **low absolute probability** of being generated by the student. Conversely, alignment does not entail exact agreement with the student's outputs, but instead requires that the corresponding tokens have **relatively higher likelihoods than other candidates** in the vocabulary.

Building on this insight, we propose that *effective reasoning trajectories should deviate from the student's own generations while remaining compatible with the student's overall generation patterns learned from prior experience*. Therefore, tokens in such trajectories are assigned low absolute probability (i.e., high surprisal) by the student model while still ranking relatively high (i.e., having low rank values) in its prediction distribution.

To validate these quantitative patterns and identify features that characterize effective learning tra-

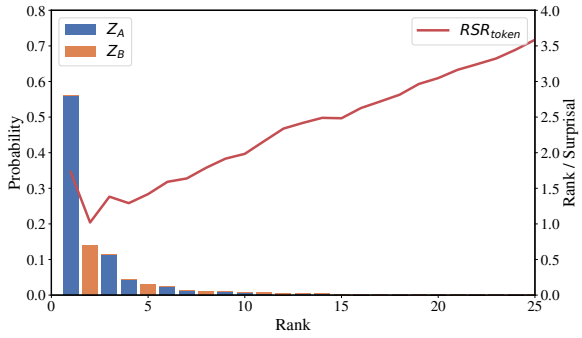

Figure 2: Simulation of the student model's token-level bimodal prediction distribution Z. Tokens with low probability yet ranked near the top (typical of the minor mode Z_B) tend to yield smaller, and thus preferred, rank-surprisal ratios.

jectories, we conduct a simulation study.

Simulation Setting We simulate the student model's token-level prediction distribution for reasoning trajectories and examine the numerical patterns exhibited by teacher trajectories that we deem effective. Specifically, we model the student's prediction distribution as a *bimodal distribution* over the vocabulary $\mathcal{V}$. The first (major) mode, denoted as Z_A, represents tokens that follow the student's dominant generation patterns, which arise from its general training data. The second (minor) mode, Z_B, occupies a small fraction of the probability mass. It represents tokens from specific generation patterns that deviate from the major mode yet remain familiar to the student model. We instantiate both Z_A and Z_B as Zipf distributions (Zipf, 1936; Mikhaylovskiy, 2025) over $\mathcal{V}$. Then the student's overall token-level prediction distribution Z is constructed as a mixture of Z_A and Z_B:

$$\begin{aligned} Z &= \pi\, Z_A + (1-\pi)\, Z_B, \\ Z_A, Z_B &\sim \text{Zipf}(\alpha), \end{aligned} \tag{3}$$

where $\pi = \frac{M_A}{M_A+M_B}$ and $M_A > M_B$.

Based on Z, we simulate four types of reasoning trajectories and examine their average token surprisal and rank: (i) X_A, sampled from Z_A, representing trajectories that closely follow the student's dominant generation patterns; (ii) X_B, sampled from Z_B, representing trajectories that deviate from the dominant mode while aligning with certain minor patterns within the student; (iii) X_C, sampled from a distribution distinct from both Z_A and Z_B, representing misaligned trajectories; and (iv) X_D, sampled from Z, representing trajectories that reflect the student's overall predictive behavior. More simulation details are provided in § A.6.

Trajectories	Prob.	Surp.	Rank	$\text{RSR}_{\text{token}}$
X_A (from Z_A, major)	0.41	1.38	2.49	1.69
X_B (from Z_B, minor)	0.10	2.73	4.31	**1.30**
X_C (from Z_C, misaligned)	0.08	4.73	11.57	2.23
X_D (from Z, mixed)	0.35	1.67	2.93	1.62

Table 3: Simulation results for different types of trajectories, reporting average token probability, surprisal, rank, and $\text{RSR}_{\text{token}}$. Effective reasoning trajectories are expected to resemble X_B, indicated by the shaded row. $\text{RSR}_{\text{token}}$ shows promise as a reliable metric for identifying effective reasoning trajectories (§ 3.4).

Simulation Results Figure 2 presents the simulated bimodal distribution Z, where effective reasoning trajectories are expected to align with the distribution of Z_B, and thus resemble X_B. As shown in Table 3, X_B exhibits a much higher average surprisal (lower probability) than X_A and X_D, while maintaining relatively low rank values (top-ranked), consistent with the quantitative patterns implied by our prior analysis. Crucially, the results suggest a promising direction for measuring data suitability, which we formalize in § 3.4.

3.4 Proposed Metric: *Rank-Surprisal Ratio*

Token-Level Metric Motivated by the complementary properties of surprisal and rank, we explore metrics that combine these two signals to jointly capture informativeness and alignment. One promising choice is the token-level ratio of rank to surprisal, which captures the relative relationship between the two signals and is denoted as $\text{RSR}_{\text{token}}$:

$$\text{RSR}_{\text{token}}(t_k) = \frac{\text{Rank}(t_k)}{\text{Surprisal}(t_k)}. \tag{4}$$

Table 3 shows that trajectories of the preferred type X_B achieve the lowest average $\text{RSR}_{\text{token}}$ (1.30), whereas the misaligned X_C achieve the highest, suggesting that this ratio may be a reliable indicator for identifying effective reasoning trajectories that balance informativeness and alignment.

From Token-Level to Trajectory-Level While $\text{RSR}_{\text{token}}$ shows encouraging behavior in simulation, directly applying this token-level ratio to assess the overall suitability of a real trajectory presents non-trivial challenges. In particular, naively averaging $\text{RSR}_{\text{token}}$ over all response tokens often leads to large and unstable values (Table 2). This instability stems from tokens that receive extremely high probabilities under certain contexts, yielding near-zero surprisal and unbounded token-level ratios due to division by near-zero values, which dominate the trajectory-level average.

A natural solution is to exclude tokens with very low surprisal when computing the average. Let $\mathcal{T}_H(\mathbf{x})$ denote the set of tokens whose surprisal lies in the top $H\%$ within a trajectory $\mathbf{x}$. We define a filtered average as

$$\text{Avg-RSR}^{\text{filter}}_{\text{token}}(\mathbf{x}) = \frac{\sum_{t_k \in \mathcal{T}_H(\mathbf{x})} \text{RSR}_{\text{token}}(t_k)}{|\mathcal{T}_H(\mathbf{x})|} \tag{5}$$

Empirically, we find that using the top 30% highest-surprisal tokens yields stronger correlation with post-training performance (Table 2). This suggests that response tokens with higher surprisal have a greater impact on student learning and should therefore be emphasized when computing the average.

Accordingly, instead of hard filtering, we adopt a surprisal-weighted average of the token-level ratios. A simple derivation shows that this weighted average is equivalent to a trajectory-level ratio between the sum of token ranks and the sum of token surprisals. For brevity, we denote $r_k = \text{Rank}(t_k)$ and $s_k = \text{Surprisal}(t_k)$:

$$\begin{aligned} \frac{\sum_k s_k \, \text{RSR}_{\text{token}}(t_k)}{\sum_k s_k} &= \frac{\sum_k s_k \, \frac{r_k}{s_k}}{\sum_k s_k} \\ &= \frac{\sum_k \text{Rank}(t_k)}{\sum_k \text{Surprisal}(t_k)} \end{aligned} \tag{6}$$

The resulting metric yields a concise form, interpretable as the ratio of the average token rank to the average token surprisal over a trajectory's response tokens.

In practice, we further observe that extremely unfamiliar tokens can attain very large rank values due to the large vocabulary size, which also leads to numerical instability. Since such tokens are effectively indistinguishable to the student, we clip rank values at a threshold r_{max} to improve stability without sacrificing meaningful information. Thus, we define our final trajectory-level metric, *Rank-Surprisal Ratio* (RSR)[2], as

$$\text{RSR}(\mathbf{x}) = \frac{\sum_k \min\big(\text{Rank}(t_k), r_{max}\big)}{\sum_k \text{Surprisal}(t_k)} \tag{7}$$

Interpretation Our metric admits a simple interpretation. The numerator, *Rank*, captures relative familiarity: lower rank values indicate that tokens in this trajectory are preferred among alternative candidates by the student and align with the model's existing behavior. The denominator, *Surprisal*, captures absolute unfamiliarity: higher surprisal indicates greater deviation from dominant patterns and provides more informative learning signals. A lower *RSR* therefore identifies trajectories that better balance alignment and informativeness, corresponding to effective reasoning supervision. § D.2 presents illustrative examples of RSR measurements.

[2]By default, RSR refers to the trajectory-level RSR in this paper, unless otherwise specified (e.g., in correlation analysis).

4 Correlation Analysis

Preliminary results in Table 2 have shown that *Rank-Surprisal Ratio* aligns well with post-training reasoning performance. To provide a more rigorous evaluation and further demonstrate the effectiveness of RSR in measuring data-student suitability, we conduct comprehensive correlation analyses.

4.1 Main Analysis

For each of the five student models and each metric, we aggregate trajectory-level suitability (or quality) scores to obtain dataset-level scores for reasoning datasets generated by eleven teacher models (§ 2.1). We then measure the correlation between these dataset-level scores and the student's reasoning performance after training on the corresponding teacher-generated datasets. We primarily report Spearman's correlation coefficient, while Pearson correlation exhibits similar trends (§ E.2). For dataset-level RSR, we adopt a weighted averaging scheme (§ A.8) similar to Eq. 6, which yields slightly higher correlation than a simple average of trajectory-level RSR. We use a clipping threshold of $r_{max} = 100$ for RSR in all subsequent experiments. Additional analysis details are provided in § A.7. Complete metric scores are reported in § F.

Compared Metrics We compare RSR against a diverse set of metrics for evaluating reasoning trajectories. These include previously discussed teacher-side indicators (e.g., teacher model performance), basic statistics such as token length, as well as probability-based metrics (e.g., average surprisal and local surprisal (Just et al., 2025)) and rank-based metrics. We also consider commonly used trajectory quality measures, such as rule-based quality scores derived from word frequency (Ye et al., 2025b), LLM-judged quality scores, and answer accuracy on verifiable questions. In addition, we include other recent student-specific data suitability metrics, including gradient-based scores (G-Norm and GRACE (Panigrahi et al., 2025)) and influence scores (Humane et al., 2025).

	Metrics	(Absolute) Spearman Correlation with Post-Training Performance					
		Qwen-3-14B	**LLaMA-3.1-8B**	**Qwen-2.5-7B**	**Qwen-3-4B**	**Qwen-2.5-3B**	**Average**
Student-Agnostic	Teacher Params	0.04	0.34	0.2	0.02	0.26	0.01
	Teacher Performance	0.49	0.34	0.13	0.23	0.03	0.23
	Avg-Token Length	0.49	0.68	0.45	0.57	0.47	0.53
	Verified Accuracy	0.54	0.43	0.25	0.35	0.10	0.33
	LLM-judged Quality	0.61	0.52	0.46	0.61	0.40	0.52
	Rule-based Quality	0.55	0.56	0.75	0.65	0.75	0.65
Student-Specific	Avg-Surprisal	0.24	0.42	0.55	0.55	0.70	0.49
	Avg-$\text{Surp}_{\text{local}}$	0.31	0.40	0.54	0.59	0.72	0.51
	Avg-Rank	0.41	0.64	0.68	0.61	0.62	0.59
	Influence Score	0.52	0.19	0.32	0.47	0.59	0.11
	G-Norm	0.44	0.54	0.51	0.57	0.70	0.55
	GRACE	0.25	0.58	0.66	0.75	0.69	0.59
	Rank-Surprisal Ratio	**0.85**	**0.85**	**0.92**	**0.82**	**0.85**	**0.86**

Table 4: Spearman correlation between each data-student suitability metric and post-training reasoning performance (average accuracy on reasoning benchmarks), reported in absolute values for different student models. "Student-agnostic" metrics are computed independently of the specific student model.

Results Table 4 shows that *RSR consistently exhibits strong correlation with post-training reasoning performance across all student models*, achieving an average Spearman correlation of 0.86 and outperforming all alternative metrics. These results indicate the effectiveness and practical value of RSR. In contrast, surprisal-based and rank-based metrics alone yield substantially weaker correlations (at most 0.59). Analysis (e.g., § 3.2) suggests that they tend to emphasize high-likelihood trajectories while insufficiently capturing informativeness. This highlights the importance of jointly modeling informativeness and alignment via the rank-surprisal ratio.

4.2 Ablation Study

The derivation of *Rank-Surprisal Ratio* involves several design components, as well as a hyperparameter r_{max}. We conduct an ablation study to examine how these choices affect the correlation strength of dataset-level RSR.

As shown in Table 5, removing either rank clipping or the surprisal-weighted averaging substantially degrades the correlation, validating the necessity of both components in our metric. In addition, the "Reduced sample size" setting estimates the dataset-level RSR using only 200 trajectories per teacher instead of the full 5,000. The comparable correlations observed under reduced sample size and alternative hyperparameter settings (e.g., $r_{max} = 500$) indicate that RSR is robust to both data scarcity and reasonable variations in r_{max}. Additional ablation results are in § E.1.

Variants	**Avg. Corr.**	Δ
Rank-Surprisal Ratio ($r_{max} = 100$)	0.856	
No rank clipping	0.700	−0.156
No weighted avg. (Avg-$\text{RSR}_{\text{token}}$)	0.391	−0.465
Filtered average (Avg-$\text{RSR}^{\text{filter}}_{\text{token}}$)	0.793	−0.064
Rank clipping: $r_{max} = 50$	0.696	−0.160
Rank clipping: $r_{max} = 500$	0.822	−0.034
Reduced sample size (200)	0.864	0.007

Table 5: Ablation study for *Rank-Surprisal Ratio*. Δ denotes the change in average correlation.

5 Practical Applications

Given the reliable data-student suitability estimation provided by *Rank-Surprisal Ratio* and its strong correlation with post-training performance, we further examine its practical value as a data selection criterion in two representative scenarios.

5.1 Trajectory Selection

Experimental Setting The trajectory selection task aims to identify the most effective reasoning trajectory from a set of candidates for a given problem or prompt. This aligns with the widespread need for data-efficient training in real-world scenarios, particularly when resources are limited. Specifically, we adopt a 33-to-1 setting, where each candidate pool contains 33 trajectories generated by 11 teacher models (3 per teacher; see § 2.1), and one trajectory is selected according to the selection criterion. This procedure is repeated for all 5,000 training problems and all student models, yielding student-specific 5,000-trajectory teacher datasets for each selection method. We then fine-tune student models on the constructed datasets and evaluate their post-training reasoning performance

Selection Methods	Qwen-3-14B					L3.1-8B	Q2.5-7B	Q3-4B	Q2.5-3B
	AIME24	AIME25	AMC23	MATH500	**Avg.**	**Avg.**	**Avg.**	**Avg.**	**Avg.**
Random	59.2	46.7	86.2	88.6	70.2	22.1	45.7	53.9	27.9
Token Length$_{max}$	61.7	51.7	87.5	84.8	71.4	27.3	45.4	51.3	27.1
Rule-based Quality$_{max}$	58.3	47.5	91.3	92.0	72.3	25.8	51.6	58.0	31.2
LLM-judged Quality$_{max}$	60.0	49.2	90.6	93.6	73.4	25.6	51.8	59.1	32.8
Surprisal$_{min}$	62.5	50.0	88.1	92.8	73.4	23.5	46.4	53.3	28.9
G-Norm$_{min}$	59.2	50.0	89.4	92.4	72.7	26.1	49.5	59.1	30.9
Rank-Surprisal Ratio$_{min}$	**67.5**	**59.2**	**93.1**	**94.6**	**78.6**	**28.5**	**53.2**	**61.4**	**34.8**

Table 6: Trajectory selection results showing post-training reasoning performance of student models trained on datasets selected by different methods (5k samples each). max and min indicate maximizing or minimizing the corresponding metric. Model names are abbreviated as **Q** for Qwen and **L** for LLaMA. Further results, including complete scores and additional GPQA evaluation, are provided in § F and § E.3.

on standard math benchmarks, consistent with previous experiments. We compare RSR against a random selection baseline and several previously discussed metrics. For metric-based methods, candidate trajectories are scored and selected by either maximizing or minimizing the corresponding trajectory-level score.

Results As shown in Table 6, datasets selected by *Rank-Surprisal Ratio* consistently achieve the best post-training reasoning performance among all selection methods across student models, demonstrating the effectiveness of RSR in identifying suitable trajectories. Moreover, the performance achieved by RSR is comparable to, and for four students even surpasses, the best performance of any single teacher for each student model (Table 1), which serves as a strong upper bound obtained via brute-force search over teacher datasets. These results underscore the practical value of RSR for selecting effective trajectories prior to training.

Complete scores are reported in § F. Additional trajectory selection results, including GPQA evaluation, No-Selection baseline, further analyses, and reduced-teacher experiments, are provided in § E.3.

5.2 Teacher Selection

Experimental Setting The teacher selection task aims to identify the most suitable teacher model for generating reasoning trajectories to train a given student model prior to distillation, reflecting a recurring challenge in real-world applications. To better capture practical constraints, we consider a low-resource setting in which generating full training data for every candidate teacher model is either costly or infeasible. Instead, we sample a small set of 200 trajectories from each candidate teacher, evaluate them using data-student suitability metrics, and select the teacher model based on the resulting dataset-level average score. We use 6 diverse teacher models (Deepseek-R1, Qwen-3-235B-Thinking, Nemotron-Super, Qwen-3-30B-Thinking, Magistral-Small, and GPT-OSS-20B) as the candidate pool, avoiding consistently well-performing teachers to ensure a non-trivial selection task.

Results As shown in Table 7, both the best and second-best teachers selected by *Rank-Surprisal Ratio* yield strong post-training reasoning performance, achieving average results close to oracle teachers and outperforming other selection methods. Notably, as also observed in our ablation study (Table 5), RSR remains effective when using only 200 trajectories per teacher for measurement, highlighting its robustness and practical value for identifying suitable teachers in low-resource settings.

6 Related Work

Knowledge Distillation Knowledge distillation is a powerful approach for transferring knowledge from large models to smaller ones (Hinton et al., 2015), and has been widely used in training LLMs (Peng et al., 2023). Prior work shows that stronger teachers do not necessarily yield better students, often due to capability gaps (Zhang et al., 2025a; Luo et al., 2025) or off-policy data (Chen et al., 2025a). Recent studies address this by better aligning teacher supervision with student behavior, achieving an implicit balance through approaches such as on-policy distillation (Agarwal et al., 2023), self-distillation (Zelikman et al., 2022; Shenfeld et al., 2026), integration with reinforcement learning (Ma et al., 2025; Zhang et al., 2025d), SFT optimization (Ye et al., 2025a; Wu et al., 2025b), teaching assistants (Mirzadeh et al., 2020; Ding et al., 2025), uncertainty-based filtering (Zhang

Selection Methods	Qwen-3-14B		LLaMA-3.1-8B		Qwen-2.5-7B		Qwen-3-4B		Qwen-2.5-3B		Avg.
	Top-1	Top-2	**Top-1**	Top-2	**Top-1**	Top-2	**Top-1**	Top-2	**Top-1**	Top-2	
Teacher Params$_{max}$	77.1	71.8	**28.1**	22.0	47.8	45.0	55.8	53.4	29.6	26.4	45.7
Token Length$_{max}$	71.8	77.1	22.0	28.1	45.0	47.8	53.4	55.8	26.4	29.6	45.7
Rule-based Quality$_{max}$	72.2	77.1	23.7	28.1	48.2	47.8	56.4	55.8	**33.0**	29.6	47.2
LLM-judged Quality$_{max}$	71.8	77.2	22.0	26.7	45.0	50.0	53.4	58.8	26.4	31.2	46.3
Surprisal$_{min}$	68.8	72.2	22.8	23.7	47.6	48.2	52.2	56.4	30.6	33.0	45.6
GRACE$_{min}$	72.2	68.8	22.8	28.1	47.6	48.3	52.2	58.8	30.6	26.4	45.6
Rank-Surprisal Ratio$_{min}$	**77.2**	77.1	26.7	28.1	**50.0**	47.8	**58.8**	55.8	31.2	30.6	**48.3**
Oracle	77.2	77.1	28.1	26.7	50.0	48.2	58.8	56.4	33.0	31.2	48.7

Table 7: Teacher selection results showing post-training reasoning performance of student models trained on data from teachers selected by different methods. Top-1 and Top-2 denote the highest- and second-highest-ranked teachers for each student. "Oracle" corresponds to the ground-truth best and second-best teachers.

et al., 2025b), and interleaved sampling (Xu et al., 2025; Peng et al., 2025). By contrast, our work explicitly quantifies the trade-off between informativeness and alignment in distillation, enabling a principled identification of effective teacher data for a given student.

SFT with Reasoning Trajectories Long CoT trajectories provide strong supervision for improving student models' reasoning performance via SFT (Wei et al., 2022; Shrestha et al., 2025). In line with findings in the general domain that high-quality data improves SFT (Zhou et al., 2023; Yang et al., 2025c), many studies focus on constructing high-quality CoT data, either by selecting better prompts (Yu et al., 2025; Yang et al., 2025b) or by filtering reasoning trajectories (Ye et al., 2025b; Jiang et al., 2025; Wu et al., 2025a; Zou et al., 2025; Chen et al., 2025b; Liu et al., 2025). Recognizing that optimal reasoning data may vary across students, recent work explores student-specific trajectory selection strategies (Panigrahi et al., 2025; Just et al., 2025; Humane et al., 2025). Our work also studies student-specific data selection, but differs from prior work by measuring data-student suitability from the perspective of informative alignment and by conducting more comprehensive evaluations across a wider range of teacher models.

7 Broader Applicability of RSR

Beyond CoT Data Although our work focuses solely on reasoning tasks, RSR is not specifically designed for reasoning trajectories (CoTs) and can be applied to general text data. Exploring its behavior and effectiveness beyond reasoning settings remains an interesting direction for future work, and we welcome efforts from the community in this direction.

Subset Selection Beyond trajectory and teacher selection, RSR may also be applicable to subset selection. Subset selection aims to identify a high-quality subset from a dataset composed of different reasoning problems and their corresponding trajectories. Unlike trajectory selection, which compares multiple trajectories for the same problem, subset selection requires cross-problem comparison of RSR values to filter samples across heterogeneous reasoning prompts. This scenario is particularly relevant when only a single trajectory is available per problem, yet improved data efficiency is still desired. Although subset selection introduces additional confounding factors due to variation across problems—making it harder to isolate trajectory-student suitability effects—RSR may still help identify relatively more effective samples across different reasoning problems. Preliminary experiments on internal data provide suggestive evidence for this possibility.

8 Conclusion

In this paper, we study data-student suitability in reasoning distillation and propose *Rank-Surprisal Ratio* (RSR), a simple metric for identifying suitable reasoning trajectories for a given student. Motivated by our analysis, RSR jointly captures a trajectory's informativeness and alignment with the student's behavior, favoring trajectories with low absolute probability but relatively high-ranked tokens. Experiments across diverse teacher-student pairs show that RSR strongly correlates with post-training performance and consistently outperforms existing metrics. We further demonstrate its effectiveness in both trajectory and teacher selection. Overall, our results highlight informative alignment as a promising direction for reasoning distillation.

Limitations

Although our work proposes an effective metric for selecting reasoning trajectories to distill student models, the performance of data selection is inherently constrained by the diversity and quality of candidate trajectories or teacher models. When none of the available teacher trajectories are well suited to a given student, the gains from selection alone may be limited. A promising direction for future work is to use our metric to guide the rewriting or synthesis of reasoning trajectories, rather than selecting from a fixed pool.

In addition, given the simple and interpretable form of the derived RSR metric, it is natural to ask whether it can be grounded in deeper theoretical principles. However, we have not yet identified a suitable theoretical framework to characterize RSR, which we leave for future investigation.

Finally, due to resource constraints, we focus primarily on mathematical reasoning tasks to conduct extensive controlled studies. Although we perform additional evaluation on GPQA to assess the generalizability of RSR, we acknowledge that the current experiments still have limitations. Specifically, we have not yet applied RSR to newly generated trajectories beyond the existing mathematical reasoning data, nor to qualitatively different domains such as code or commonsense reasoning. Systematic cross-domain studies would be valuable, but they require large-scale trajectory regeneration and retraining, entailing substantial computational cost. Therefore, we focus on the current setting (following prior work) and leave such systematic extensions for future work.

Acknowledgements

The authors wish to thank the anonymous reviewers for their helpful comments. This work was partially funded by National Natural Science Foundation of China (No. 62476061, 62521004, 62576106, 62376061).

References

Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, and 1 others. 2025. Phi-4-reasoning technical report. *arXiv preprint arXiv:2504.21318*.

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stańczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. 2023. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*.

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, and 1 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, and 1 others. 2025. Llama-nemotron: Efficient reasoning models. *arXiv preprint arXiv:2505.00949*.

Howard Chen, Noam Razin, Karthik Narasimhan, and Danqi Chen. 2025a. Retaining by doing: The role of on-policy data in mitigating forgetting. *CoRR*, abs/2510.18874.

Xinghao Chen, Zhijing Sun, Wenjin Guo, Miaoran Zhang, Yanjun Chen, Yirong Sun, Hui Su, Yijie Pan, Dietrich Klakow, Wenjie Li, and Xiaoyu Shen. 2025b. Unveiling the key factors for distilling chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 15094–15119. Association for Computational Linguistics.

Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4.

Tri Dao. 2024. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*.

Dongyi Ding, Tiannan Wang, Chenghao Zhu, Meiling Tao, Yuchen Eleanor Jiang, and Wangchunshu Zhou. 2025. Micota: Bridging the learnability gap with intermediate cot and teacher assistants. *ArXiv*, abs/2507.01887.

Lishui Fan, Yu Zhang, Mouxiang Chen, and Zhongxin Liu. 2025. Posterior-grpo: Rewarding reasoning processes in code generation. *CoRR*, abs/2508.05170.

Roger B. Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, Evan Hubinger, Kamile Lukosiute, Karina Nguyen, Nicholas Joseph, Sam McCandlish, Jared Kaplan, and Samuel R. Bowman. 2023. Studying large language model generalization with influence functions. *CoRR*, abs/2308.03296.

Etash Kumar Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer, Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline

Choi, Niklas Muennighoff, Shiye Su, and 31 others. 2025. Openthoughts: Data recipes for reasoning models. *CoRR*, abs/2506.04178.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nat.*, 645(8081):633–638.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Language Technologies 2001: The Second Meeting of the North American Chapter of the Association for Computational Linguistics, NAACL 2001, Pittsburgh, PA, USA, June 2-7, 2001*. The Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531.

Prateek Humane, Paolo Cudrano, Daniel Z. Kaplan, Matteo Matteucci, Supriyo Chakraborty, and Irina Rish. 2025. Influence functions for efficient data selection in reasoning. *CoRR*, abs/2510.06108.

Gangwei Jiang, Yahui Liu, Zhaoyi Li, Qi Wang, Fuzheng Zhang, Linqi Song, Ying Wei, and Defu Lian. 2025. What makes a good reasoning chain? uncovering structural patterns in long chain-of-thought reasoning. *CoRR*, abs/2505.22148.

Hoang Anh Just, Myeongseob Ko, and Ruoxi Jia. 2025. Distilling reasoning into student llms: Local naturalness for selecting teacher data. *CoRR*, abs/2510.03988.

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

Dacheng Li, Shiyi Cao, Tyler Griggs, Shu Liu, Xiangxi Mo, Eric Tang, Sumanth Hegde, Kourosh Hakhamaneshi, Shishir G. Patil, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. 2025a. Llms can easily learn to reason from demonstrations structure, not content, is what matters! *CoRR*, abs/2502.07374.

Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Huang, Kashif Rasul, Longhui Yu, Albert Q Jiang, Ziju Shen, and 1 others. 2024. Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions. *Hugging Face repository*, 13(9):9.

Yuetai Li, Xiang Yue, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, Bhaskar Ramasubramanian, and Radha Poovendran. 2025b. Small models struggle to learn from strong reasoners. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 25366–25394. Association for Computational Linguistics.

Jinrui Liu, Jeff Wu, Xuanguang Pan, Gavin Cheung, Shuai Ma, and Chongyang Tao. 2025. Air: Post-training data selection for reasoning via attention head influence. *arXiv preprint arXiv:2512.13279*.

Renjie Luo, Jiaxi Li, Chen Huang, and Wei Lu. 2025. Through the valley: Path to effective long cot training for small language models. *CoRR*, abs/2506.07712.

Lu Ma, Hao Liang, Meiyi Qiang, Lexiang Tang, Xiaochen Ma, Zhen Hao Wong, Junbo Niu, Chengyu Shen, Runming He, Bin Cui, and Wentao Zhang. 2025. Learning what reinforcement learning can't: Interleaved online fine-tuning for hardest questions. *CoRR*, abs/2506.07527.

Nikolay Mikhaylovskiy. 2025. Zipf's and heaps' laws for tokens and llm-generated texts. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15469–15481.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. 2020. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5191–5198.

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

Abhishek Panigrahi, Bingbin Liu, Sadhika Malladi, Sham M. Kakade, and Surbhi Goel. 2025. In good graces: Principled teacher selection for knowledge distillation. *CoRR*, abs/2511.02833.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with GPT-4. *CoRR*, abs/2304.03277.

Jingyu Peng, Maolin Wang, Hengyi Cai, Yuchen Li, Kai Zhang, Shuaiqiang Wang, Dawei Yin, and Xiangyu Zhao. 2025. Adaswitch: Adaptive switching generation for knowledge distillation. *ArXiv*, abs/2510.07842.

Abhilasha Ravichander, Jillian Fisher, Taylor Sorensen, Ximing Lu, Maria Antoniak, Bill Yuchen Lin, Niloofar Mireshghallah, Chandra Bhagavatula, and Yejin Choi. 2025. Information-guided identification of training data imprint in (proprietary) large language

models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 1962–1978. Association for Computational Linguistics.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. 2026. Self-distillation enables continual learning. *CoRR*, abs/2601.19897.

Safal Shrestha, Minwu Kim, Aadim Nepal, Anubhav Shrestha, and Keith Ross. 2025. Warm up before you train: Unlocking general reasoning in resource-constrained settings. *CoRR*, abs/2505.13718.

Lev Semenovich Vygotsky and Michael Cole. 1978. *Mind in society: Development of higher psychological processes*. Harvard university press.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Xiaojun Wu, Xiaoguang Jiang, Huiyang Li, Jucai Zhai, Dengfeng Liu, Qiaobo Hao, Huang Liu, Zhiguo Yang, Ji Xie, Ninglun Gu, Jin Yang, Kailai Zhang, Yelun Bao, and Jun Wang. 2025a. Beyond scaling law: A data-efficient distillation framework for reasoning. *CoRR*, abs/2508.09883.

Yongliang Wu, Yizhou Zhou, Zhou Ziheng, Yingzhe Peng, Xinyu Ye, Xinting Hu, Wenbo Zhu, Lu Qi, Ming-Hsuan Yang, and Xu Yang. 2025b. On the generalization of SFT: A reinforcement learning perspective with reward rectification. *CoRR*, abs/2508.05629.

Wenda Xu, Rujun Han, Zifeng Wang, Long T. Le, Dhruv Madeka, Lei Li, William Yang Wang, Rishabh Agarwal, Chen-Yu Lee, and Tomas Pfister. 2025. Speculative knowledge distillation: Bridging the teacher-student gap through interleaved sampling. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 40 others. 2025a. Qwen3 technical report. *CoRR*, abs/2505.09388.

Cehao Yang, Xueyuan Lin, Chengjin Xu, Xuhui Jiang, Xiaojun Wu, Honghao Liu, Hui Xiong, and Jian Guo. 2025b. Select2reason: Efficient instruction-tuning data selection for long-cot reasoning. *CoRR*, abs/2505.17266.

Yuming Yang, Yang Nan, Junjie Ye, Shihan Dou, Xiao Wang, Shuo Li, Huijie Lv, Tao Gui, Qi Zhang, and Xuanjing Huang. 2025c. Measuring data diversity for instruction tuning: A systematic analysis and A reliable metric. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 18530–18549. Association for Computational Linguistics.

Junjie Ye, Yuming Yang, Yang Nan, Shuo Li, Qi Zhang, Tao Gui, Xuan-Jing Huang, Peng Wang, Zhongchao Shi, and Jianping Fan. 2025a. Analyzing the effects of supervised fine-tuning on model knowledge from token and parameter levels. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 471–513.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025b. LIMO: less is more for reasoning. *CoRR*, abs/2502.03387.

Qianjin Yu, Keyu Wu, Zihan Chen, Chushu Zhang, Manlin Mei, Lingjun Huang, Fang Tan, Yongsheng Du, Kunlin Liu, and Yurui Zhu. 2025. Rethinking the generation of high-quality cot data from the perspective of llm-adaptive question difficulty grading. *CoRR*, abs/2504.11919.

Jerrold H Zar. 2005. Spearman rank correlation. *Encyclopedia of biostatistics*, 7.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. Star: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Chen Zhang, Qiuchi Li, Dawei Song, Zheyu Ye, Yan Gao, and Yao Hu. 2025a. Towards the law of capacity gap in distilling language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 22504–22528. Association for Computational Linguistics.

Dylan Zhang, Qirun Dai, and Hao Peng. 2025b. The best instruction-tuning data are those that fit. *CoRR*, abs/2502.04194.

Qiyuan Zhang, Fuyuan Lyu, Zexu Sun, Lei Wang, Weixu Zhang, Zhihan Guo, Yufei Wang, Irwin King, Xue Liu, and Chen Ma. 2025c. What, how, where, and how well? A survey on test-time scaling in large language models. *CoRR*, abs/2503.24235.

Xuechen Zhang, Zijian Huang, Yingcong Li, Chenshun Ni, Jiasi Chen, and Samet Oymak. 2025d. BREAD: branched rollouts from expert anchors bridge SFT & RL for reasoning. *CoRR*, abs/2506.17211.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan 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. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

George Kingsley Zipf. 1936. *The psycho-biology of language: An introduction to dynamic philology*. Routledge.

Jiaru Zou, Ling Yang, Jingwen Gu, Jiahao Qiu, Ke Shen, Jingrui He, and Mengdi Wang. 2025. Reasonflux-prm: Trajectory-aware prms for long chain-of-thought reasoning in llms. *CoRR*, abs/2506.18896.

A Details of Experiments

A.1 Determining Training Problem Set

All teacher trajectory datasets are constructed using a fixed problem set to avoid confounding effects from variations in problem composition, enabling a more controlled study. We focus on mathematical reasoning and curate 5,000 math problems from the widely used NuminaMath dataset (Li et al., 2024). Following prior work (Sky-T1) (Li et al., 2025a), we apply preprocessing steps such as difficulty filtering to ensure the problem quality.

Specifically, our training dataset consists of problems drawn from the MATH, AIME/AMC, and Olympiads. Sky-T1 provides a difficulty-labeled version of NuminaMath in which each problem is assigned an integer difficulty score from 0 to 9. Using this scale, we randomly sampled 1,667 problems from the MATH subset with difficulty above three, 1,667 problems from the Olympiads subset with difficulty above eight, and 1,666 problems from the AIME/AMC subset, yielding a balanced training set of 5,000 problems.

We fix the training set size at 5,000 problems for three reasons. First, prior studies (Ye et al., 2025b; Muennighoff et al., 2025) have shown that strong reasoning capabilities can be learned from training sets of around 1,000 problems. Second, high-quality reasoning problems are relatively scarce, making further scaling often unrealistic in practice. Third, our distillation study incurs quadratic computational costs across teacher-student pairs. Considering these factors, we find 5,000 problems to offer a good trade-off between representativeness and computational feasibility.

A.2 Teacher Trajectory Generation

For each teacher model, we generate reasoning trajectories for 5,000 problems over three independent runs using vLLM (Kwon et al., 2023) under a maximum generation budget of 31,000 tokens. Each independent run (rollout) produces a dataset of 5,000 problem-trajectory pairs, yielding $11 \times 3 = 33$ datasets in total.

We adopt the officially recommended chat template and model sampling hyperparameters; for example, Qwen models use temperature=0.6, top_p=0.95, top_k=20, and min_p=0. Each problem is appended with the instruction: "Return your final response within \boxed{}." If a sampled trajectory exceeds the token budget, we resample up to 10 times; if all attempts still exceed the budget, we truncate the final trajectory. Across all teacher models, final truncation rates are below 1%, which helps preserve the quality of training trajectories even for teachers that tend to produce long outputs.

The resulting training dataset of teacher trajectories is formatted as follows: for each sample, we use the original problem statement as the user prompt and a single corresponding teacher-generated trajectory as the assistant response, with a fixed system prompt applied to all instances: "Please reason step by step, and put your final answer within \boxed{}." to each problem.

We release all 33 generated trajectory datasets along with RSR-selected subsets (§ 5.1) on Hugging Face.[3]

A.3 Teacher Models and Student Models

Teacher Models
DeepSeek-R1-0528
Qwen-3-235B-Thinking-2507
GPT-OSS-120B (high)
LLaMA-3.3-Nemotron-Super-49B-v1.5
QwQ-32B
Qwen-3-30B-Thinking-2507
Magistral-Small-2506
GPT-OSS-20B (high)
Phi-4-Reasoning-Plus
Qwen-3-8B (thinking)
Qwen-3-4B-Thinking-2507

Table 8: List of teacher models used in our experiments.

As discussed earlier, our experiments adopt a more diverse set of teacher models than prior work. Specifically, we consider the teacher variants shown in Table 8. This set includes several recently emerged reasoning models, extending beyond the DeepSeek-R1 and QwQ teachers commonly used in prior studies.

These teachers produce trajectories that vary in length, informational content, and style. For example, trajectories generated by GPT-OSS models tend to be concise, whereas those from DeepSeek-R1 are generally more detailed. Specific trajectory examples from different teachers are available in our released datasets on Hugging Face and in our case study (§ D.2).

For the student models, we select five open-source pretrained base models from the Qwen and LLaMA families: Qwen-3-14B, LLaMA-3.1-8B, Qwen-2.5-7B, Qwen-3-4B, and Qwen-2.5-3B. We

[3] https://huggingface.co/datasets/Umean/RSR_data

use base models instead of chat models as students, following the training practice of recent reasoning models. This choice ensures a clean starting point with greater potential (Yang et al., 2025a), allows us to observe more substantial training effects, and avoids stylistic mismatch and overlapping supervision between instruction-tuning data and reasoning trajectories.

A.4 Benchmark Evaluation

We use vLLM together with the Math-Verify package to evaluate post-trained models on mathematics benchmarks. Our evaluated benchmarks, AIME'25, AIME'24, AMC'23, and MATH500, are widely used and span varying difficulty levels, assessing mathematical reasoning and multi-step problem-solving across diverse domains. We additionally conduct evaluation on GPQA-Diamond beyond mathematics, as described in § E.3.2. We adopt the Acc@4 metric as the final score, which averages results over four independent evaluations per problem. Evaluating a single model checkpoint typically takes around half an hour using 8 H200 GPUs.

During inference, we use a temperature of 0.6, a top_p of 0.95, and top_k set to -1. The maximum generation length is set to 32,768 tokens, consistent with the maximum sequence length used during fine-tuning. Responses that exceed this limit are truncated. This differs from trajectory generation, where we resample multiple times to avoid truncation, as here we aim to evaluate the model's reasoning ability under a fixed context-length constraint. Under this setting, the comparable performance of the Qwen-3-235B and Qwen-3-30B in Table 1 may be attributable to truncation effects, which we consider a reasonable outcome given the imposed length limit.

A.5 Details of Model Fine-Tuning

Models	Learning Rate	Batch Size	Epoch
Qwen-3-14B	2.0E-05	64	8
LLaMA-3.1-8B	2.0E-05	64	10
Qwen-2.5-7B	2.0E-05	64	10
Qwen-3-4B	2.0E-05	64	10
Qwen-2.5-3B	5.0E-05	64	10

Table 9: Training hyperparameters for different student models.

We perform SFT on reasoning trajectory datasets using the LLaMA-Factory (Zheng et al., 2024) framework and FlashAttention-2 (Dao, 2024). For different student models, we conduct grid search for best set of hyperparameters. The final setting is reported in the Table 9. During SFT, all student models use a maximum sequence length of 32,768.

Most experiments are conducted on NVIDIA H200 GPUs. A single SFT experiment using the 7B model takes around 7 hours on 8 H200 GPUs for training.

For trajectory selection experiments in § 5.1, we fine-tune each selected dataset with three random seeds and report performance averaged over these runs. For distillation experiments in § 2.1, however, each teacher already yields three independent trajectory datasets, so we do not further vary random seeds for fine-tuning.

A.6 Details of Simulation Study

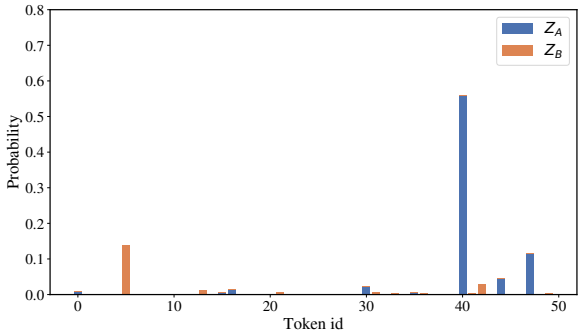

Figure 3: Simulation of the student model's token-level prediction distribution Z, a mixture of Zipf distributions Z_A and Z_B.

In practice, we simulate Z by sampling M_A tokens from Z_A and M_B tokens from Z_B. We use $M_A = 1{,}000{,}000$, $M_B = 250{,}000$, $|\mathcal{V}| = 50$, and for each simulated dataset, we sample 10,000 tokens to compute average metrics. Based on a preliminary fit to reasoning data, we set the Zipf exponent to $\alpha = 2.3$.

Figure 3 depicts the distributions of Z_A and Z_B over the vocabulary, with their mixture representing the simulated token-level prediction distribution Z of the student model. Figure 2 is derived from this by ranking the tokens from higher to lower probability.

A.7 Details of Correlation Analysis

We use the Spearman correlation coefficient (Zar, 2005) because our analysis focuses on monotonic consistency rather than strict linear relationships between metrics and post-training performance. We

also report Pearson correlation (Cohen et al., 2009) results in § E.2.

We report the absolute values of Spearman correlation coefficients in Table 4. For the "Average" values, however, we first average the correlation coefficients across student models and then take the absolute value. We consider this aggregation more appropriate, as correlations with opposite signs across different student models should offset each other; consequently, this averaged value can be lower than the result obtained by averaging absolute correlations.

Since post-training performance is computed by averaging three generation runs per teacher (§ 2.1), we likewise average the dataset-level metric over the three trajectory datasets generated by the same teacher to obtain its final score. We observe that our datset-level metric varies marginally across different datasets generated by the same teacher.

A.8 Definition of Dataset-level RSR

For dataset-level RSR, we adopt a surprisal-weighted averaging scheme analogous to the trajectory-level weighted average (§ 3.4). This design aims to mitigate numerical instability caused by trajectories with disproportionately small average surprisal and to emphasize trajectories with larger average surprisal during dataset-level aggregation. We now formally define the dataset-level RSR used in the correlation analysis (§ 4.1, 4.2).

Let $\mathbf{X} = \{\mathbf{x}_j\}_{j=1}^{|\mathbf{X}|}$ denote a dataset of trajectories, where each trajectory $\mathbf{x}_j$ contains response tokens $\{t_{j,k}\}_{k=1}^{|\mathbf{x}_j|}$. For brevity, we define the trajectory-level average clipped rank and average surprisal as

$$\begin{aligned}\overline{r}(\mathbf{x}_j) &= \frac{1}{|\mathbf{x}_j|}\sum_{k=1}^{|\mathbf{x}_j|}\min\big(\mathrm{Rank}(t_{j,k}), r_{max}\big),\\ \overline{s}(\mathbf{x}_j) &= \frac{1}{|\mathbf{x}_j|}\sum_{k=1}^{|\mathbf{x}_j|}\mathrm{Surprisal}(t_{j,k}).\end{aligned} \tag{8}$$

Accordingly, the trajectory-level RSR can be equivalently written as $\mathrm{RSR}(\mathbf{x}_j) = \frac{\overline{r}(\mathbf{x}_j)}{\overline{s}(\mathbf{x}_j)}$.

To obtain the dataset-level RSR, we apply a weighted averaging scheme analogous to Eq. 6, using the trajectory-level average surprisal $\overline{s}(\mathbf{x}_j)$ as the weight:

$$\begin{aligned}\mathrm{RSR}_{\text{dataset}}(\mathbf{X}) &= \frac{\sum_{j=1}^{|\mathbf{X}|}\overline{s}(\mathbf{x}_j)\,\mathrm{RSR}(\mathbf{x}_j)}{\sum_{j=1}^{|\mathbf{X}|}\overline{s}(\mathbf{x}_j)}\\ &= \frac{\sum_j \overline{s}(\mathbf{x}_j)\,\frac{\overline{r}(\mathbf{x}_j)}{\overline{s}(\mathbf{x}_j)}}{\sum_j \overline{s}(\mathbf{x}_j)}\\ &= \frac{\sum_j \overline{r}(\mathbf{x}_j)}{\sum_j \overline{s}(\mathbf{x}_j)}.\end{aligned} \tag{9}$$

The resulting dataset-level metric admits a concise form, which can be interpreted as the ratio of the summed trajectory-level average clipped ranks to the summed trajectory-level average surprisals.

We compare the correlation strength of the simple dataset-level average of trajectory-level RSR, i.e., $\frac{1}{|\mathbf{X}|}\sum_{j=1}^{|\mathbf{X}|}\mathrm{RSR}(\mathbf{x}_j)$, with that of the dataset-level RSR computed via surprisal-weighted averaging (Eq. 9) in the additional ablation study (§ E.1).

B Details of Compared Metrics

B.1 LLM-judged Quality

We use Qwen-3-32B-Instruct in a Non-Thinking configuration as an automatic judge to evaluate reasoning trajectories. Under a fixed evaluation prompt, the judge produces a structured assessment at two levels. First, it assigns dimension-wise scores over five criteria—Factual Accuracy, Logical Rigor, Solution Completeness, Reasoning Efficiency, and Presentation Quality—each accompanied by a brief justification. Second, it aggregates the dimensional assessments into an overall score in $[0, 1]$ together with a concise rationale. For dataset-level comparison, we report the mean of the trajectory-level overall scores. The complete evaluation prompt is given in Table 30, adapted from (Fan et al., 2025).

B.2 Rule-based Quality

We implement the rule-based criterion used for filtering the LIMO Dataset (Ye et al., 2025b). Each response is scored using the following weighted indicators: *Elaborated reasoning (30%)*: total word length. *Self-Verification (20%)*: frequency of "check" and "verify". *Exploratory Approach (25%)*: frequency of "perhaps" and "might". *Adaptive Granularity (25%)*: frequency of "therefore" and "since".

To ensure fair comparison across responses of varying lengths, we compute relative keyword frequencies by normalizing absolute keyword counts

with respect to the total word count. To account for differences in scale across criteria, we then independently standardize each criterion's scores into z-scores, which empirically improves correlation.

B.3 Influence Score

We adopt the Influence Score method based on the second-order approximation of influence functions (Humane et al., 2025). The core idea is to model an infinitesimal up-weighting of a training sample as a local perturbation to the optimization objective, thereby estimating the resulting marginal change in the evaluation-set loss. Within this framework, a higher influence score indicates that the gradient direction induced by the training sample is more aligned with minimizing the evaluation loss in the vicinity of the converged parameters.

We implement this baseline using the Kronfluence framework. To make the computation tractable for LLMs, we employ the EK-FAC strategy to approximate the Hessian matrix (Grosse et al., 2023). Specifically, we first precompute the EK-FAC factors on a reference model. Since we do not know which trajectory is better a priori, we use the model trained with randomly selected trajectories per problem as the reference model, corresponding to the "Random" variant in Table 6. Then, we compute the pairwise influence scores between the training samples and the evaluation dataset. To optimize for memory and computational efficiency, we apply a low-rank approximation to the query gradients (rank $= 4$) and utilize `bfloat16` precision for the Inverse Hessian-Vector Product calculations. Finally, we average these pairwise scores across the evaluation set to obtain the sample-level (trajectory-level) influence score $s(d)$, which is used as the baseline for evaluating data suitability.

B.4 G-Norm

G-Norm (Panigrahi et al., 2025) assesses the suitability of generated data for a student model by measuring the magnitude of the student's local gradient signal induced by the data. Formally, G-Norm calculates the magnitude of the loss gradient with respect to the student model's parameters. To address the computational constraints associated with high-dimensional parameter spaces, the method utilizes random projection for dimensionality reduction. For each reasoning trajectory $\mathbf{x}$, we compute the gradient of the loss derived from the student model and apply length normalization to eliminate bias towards longer sequences. These gradients are then projected onto a lower-dimensional subspace via a fixed random matrix, followed by the computation of their L_2 norms.

B.5 GRACE

We adopt GRACE (Panigrahi et al., 2025) as a baseline metric for further evaluating the generated reasoning trajectory dataset from a gradient-based perspective. GRACE characterizes the geometry of the optimization landscape by analyzing the spectral structure of the student model gradients. This approach allows for a holistic assessment of how effectively the generated data spans the parameter space required for model optimization.

The calculation proceeds by first projecting the high-dimensional gradients onto a lower-dimensional subspace via a fixed random matrix to ensure computational feasibility. To address the empirical tendency of gradient norms to diminish in longer sequences, the method rescales the projected vectors logarithmically based on the response length. GRACE then computes the metric using a cross-validation strategy where the dataset is divided into multiple partitions. The gradients from a held-out partition are weighted by the regularized inverse covariance matrix estimated from the remaining data. This process effectively quantifies the expected squared norm of the gradients after whitening them with the estimated spectrum of the distribution.

Because GRACE's cross-validation strategy yields a dataset-level metric rather than a per-sample score, we apply GRACE for teacher selection but not for trajectory-level selection.

B.6 Others

Here we give formal definitions for Avg-Surprisal and Avg-Surp$_{\text{local}}$ (§ 3.2). Given a trajectory $x = (t_1, \ldots, t_T)$,

$$\text{Avg\_Surprisal}(\mathbf{x}) = \frac{1}{T}\sum_{k=1}^{T} -\log p_\theta(t_k \mid \mathbf{c}_k) \tag{10}$$

$$\text{Avg\_Surp}_{\text{local}}(\mathbf{x}) = \frac{1}{T}\sum_{k=1}^{T} -\log p_\theta(t_k \mid \mathbf{c}_k^{\text{local}}) \tag{11}$$

C Computational Cost of RSR

The computation of *Rank-Surprisal Ratio* requires only a single forward pass through the student

model, during which we collect each token's surprisal and clipped rank from the model logits. For a single trajectory, computing token-level surprisals and ranks has a worst-case time complexity of $\mathcal{O}(TV)$, where T denotes the number of response tokens and V is the vocabulary size. This computation does not introduce additional complexity beyond that inherent in the dimensionality of the model logits. Moreover, since RSR only depends on clipped ranks, the rank computation can be efficiently implemented via top-k selection rather than full-vocabulary comparisons (see our code on GitHub), reducing practical runtime and GPU memory usage.

In practice, computing RSR over the 5,000-trajectory dataset with a context length of 32,768 using a 7B model typically takes under one hour on a single H200 GPU with FlashAttention-2 enabled, which is significantly cheaper than SFT.

We also conduct a direct wall-clock time comparison for computing RSR, G-Norm, Influence Score, and LLM-judged Quality on a fixed dataset of 5,000 DeepSeek-R1 trajectories using the LLaMA-3.1-8B student model. As shown in Table 10, RSR incurs significantly lower total GPU-hours than the other metrics. This likely stems from the fact that RSR requires only a single forward pass through the student model, whereas the other methods involve gradient computations or text generation.

Metrics	#GPU	Hours	Total GPU-hours
RSR	1	1	1
G-Norm	2	3	6
Influence Score	8	4	32
LLM-judged Quality	8	1.5	12

Table 10: Computational cost comparison of different metrics on a fixed set of 5,000 DeepSeek-R1 trajectories using the LLaMA-3.1-8B student model, evaluated on H200 GPUs.

D What Does RSR Prefer? Statistical Analysis and Case Study

D.1 Statistical Analysis of RSR-Selected Trajectories

We compare the statistics of the RSR-selected trajectories with the aggregate statistics (mean/-max/min) across source teacher datasets. From Table 11, the RSR-selected trajectories exhibit moderate token length and verified accuracy. This suggests that RSR is not driven solely by length or accuracy, but instead follows its own evaluation criterion. Additional analysis of datasets selected by RSR can be found in § E.3.3.

Datasets	Avg. Token Length	Verified Accuracy
RSR Selected	9845	81.19
Mean over Teachers	8912	81.58
Max.	13145	85.9
Min.	2552	77.6

Table 11: Statistics of RSR-selected trajectories and source teacher trajectories in terms of token length and verified accuracy. Qwen-2.5-7B is used as the student model. RSR follows its own evaluation criterion, independent of length or accuracy.

D.2 Token-Level Case Study

To better understand how *Rank-Surprisal Ratio* operates in practice and what kinds of trajectories it prefers, we conduct a token-level case study comparing trajectories generated by different teacher models.

Table 12 presents representative text segments from trajectories generated by GPT-OSS-20B, Nemotron-Super-49B-v1.5, and QwQ-32B, together with their token-level rank, surprisal, and RSR values as measured by the Qwen-2.5-7B student, highlighting distinct statistical patterns across teacher models. Unlike Eq. 4, we compute token-level RSR with rank clipping at $r_{\max} = 100$, consistent with trajectory-level RSR, to improve numerical stability.[4]

As shown, tokens in trajectories from GPT-OSS-20B often exhibit excessively high rank values, suggesting that these trajectories are relatively unfamiliar to the student model and misaligned with the student's current behavior. This may stem from the fast-paced reasoning flow and uncommon phrasing in GPT-OSS-20B trajectories, which make them harder for the student to follow. Meanwhile, tokens from Nemotron-Super frequently have very low surprisals, indicating that these trajectories closely resemble the student's own generations and therefore provide limited informativeness. This may be attributed to the relatively conventional reasoning flow and limited analytical depth of Nemotron-Super's trajectories. As a result, tokens from both models tend to receive high token-level RSR values, implying less effective supervision for student training. By comparison, trajectories from QwQ-32B exhibit a moderate level of comprehensibility

[4]We denote this variant as $\text{RSR}^{*}_{\text{token}}$, and all token-level RSR values in this section refer to it.

Token-level Case Study

Trajectories from GPT-OSS-20B: (High Rank Values, High RSR)

...	Check	no	extra	trick	:	cross	product	magnitude
	Rank: 153	Rank: 504	Rank: 18	Rank: 1414	Rank: 13	Rank: 20	Rank: 1	Rank: 6
	Surprisal: 7.28	Surprisal: 9.62	Surprisal: 4.91	Surprisal: 10.69	Surprisal: 4.50	Surprisal: 4.78	Surprisal: 0.24	Surprisal: 3.84
	RSR^*_{token}: 13.70	RSR^*_{token}: 10.42	RSR^*_{token}: 3.66	RSR^*_{token}: 9.35	RSR^*_{token}: 2.89	RSR^*_{token}: 4.18	RSR^*_{token}: 4.26	RSR^*_{token}: 1.56
?	Not	needed	.	Good	.	Thus	final	response
Rank: 27	Rank: 23	Rank: 1	Rank: 1	Rank: 214	Rank: 2	Rank: 10	Rank: 2	Rank: 57
Surprisal: 5.66	Surprisal: 5.03	Surprisal: 0.91	Surprisal: 1.59	Surprisal: 7.97	Surprisal: 2.70	Surprisal: 4.44	Surprisal: 2.09	Surprisal: 8.00
RSR^*_{token}: 4.78	RSR^*_{token}: 4.57	RSR^*_{token}: 1.09	RSR^*_{token}: 0.63	RSR^*_{token}: 12.50	RSR^*_{token}: 0.74	RSR^*_{token}: 2.25	RSR^*_{token}: 0.96	RSR^*_{token}: 7.14
box	...							
Rank: 24								
Surprisal: 5.91								
RSR^*_{token}: 4.07								

Trajectories from Nemotron-Super-49B-v1.5: (Low Surprisal, High RSR)

...	But	the	problem	doesn	’t	mention	the	height
	Rank: 2	Rank: 1	Rank: 1	Rank: 1	Rank: 1	Rank: 1	Rank: 1	Rank: 1
	Surprisal: 1.73	Surprisal: 0.51	Surprisal: 0.04	Surprisal: 0.17	Surprisal: 0.00	Surprisal: 0.22	Surprisal: 0.29	Surprisal: 0.12
	RSR^*_{token}: 1.16	RSR^*_{token}: 1.95	RSR^*_{token}: 25.64	RSR^*_{token}: 6.02	RSR^*_{token}: >100	RSR^*_{token}: 4.61	RSR^*_{token}: 3.41	RSR^*_{token}: 8.55
of	the	fountain	,	so	maybe	it	’s	safe
Rank: 1	Rank: 1	Rank: 1	Rank: 1	Rank: 1	Rank: 1	Rank: 2	Rank: 1	Rank: 4
Surprisal: 0.13	Surprisal: 0.00	Surprisal: 0.00	Surprisal: 0.21	Surprisal: 0.17	Surprisal: 0.65	Surprisal: 1.45	Surprisal: 0.13	Surprisal: 3.03
RSR^*_{token}: 7.46	RSR^*_{token}: >100	RSR^*_{token}: >100	RSR^*_{token}: 4.72	RSR^*_{token}: 5.99	RSR^*_{token}: 1.54	RSR^*_{token}: 1.38	RSR^*_{token}: 7.87	RSR^*_{token}: 1.32
to	assume	that	the	pl	anks	are	placed	in
Rank: 1	Rank: 1	Rank: 2	Rank: 1	Rank: 1	Rank: 1	Rank: 1	Rank: 2	Rank: 1
Surprisal: 0.00	Surprisal: 0.01	Surprisal: 1.05	Surprisal: 0.03	Surprisal: 0.04	Surprisal: 0.00	Surprisal: 0.02	Surprisal: 1.40	Surprisal: 0.16
RSR^*_{token}: >100	RSR^*_{token}: 166.67	RSR^*_{token}: 1.90	RSR^*_{token}: 37.04	RSR^*_{token}: 27.78	RSR^*_{token}: >100	RSR^*_{token}: 62.50	RSR^*_{token}: 1.43	RSR^*_{token}: 6.10
the	plane	of	the	circular	base	...		
Rank: 1	Rank: 1	Rank: 1	Rank: 1	Rank: 3	Rank: 1			
Surprisal: 0.07	Surprisal: 0.02	Surprisal: 0.06	Surprisal: 0.00	Surprisal: 3.73	Surprisal: 0.00			
RSR^*_{token}: 14.71	RSR^*_{token}: 47.62	RSR^*_{token}: 18.18	RSR^*_{token}: >100	RSR^*_{token}: 0.80	RSR^*_{token}: >100			

Trajectories from QwQ-32B: (Low RSR, Thus Preferred)

...	Let	me	think	of	the	data	as	a	list
	Rank: 4	Rank: 1	Rank: 1	Rank: 5	Rank: 2	Rank: 2	Rank: 1	Rank: 1	Rank: 4
	Surprisal: 2.70	Surprisal: 0.35	Surprisal: 1.63	Surprisal: 2.95	Surprisal: 1.34	Surprisal: 1.94	Surprisal: 0.81	Surprisal: 0.73	Surprisal: 2.88
	RSR^*_{token}: 1.48	RSR^*_{token}: 2.82	RSR^*_{token}: 0.61	RSR^*_{token}: 1.69	RSR^*_{token}: 1.49	RSR^*_{token}: 1.03	RSR^*_{token}: 1.23	RSR^*_{token}: 1.36	RSR^*_{token}: 1.39
.	Let	me	try	to	write	out	the	positions	
Rank: 2	Rank: 1	Rank: 1	Rank: 2	Rank: 1	Rank: 6	Rank: 2	Rank: 1	Rank: 21	
Surprisal: 1.87	Surprisal: 1.68	Surprisal: 0.64	Surprisal: 2.00	Surprisal: 0.98	Surprisal: 3.23	Surprisal: 1.15	Surprisal: 0.27	Surprisal: 5.84	
RSR^*_{token}: 1.07	RSR^*_{token}: 0.60	RSR^*_{token}: 1.56	RSR^*_{token}: 1.00	RSR^*_{token}: 1.02	RSR^*_{token}: 1.86	RSR^*_{token}: 1.74	RSR^*_{token}: 3.70	RSR^*_{token}: 3.60	
for	␣	1	0	students	?	Wait	...		
Rank: 3	Rank: 1	Rank: 1	Rank: 1	Rank: 2	Rank: 425	Rank: 6			
Surprisal: 2.42	Surprisal: 1.05	Surprisal: 0.08	Surprisal: 0.00	Surprisal: 5.25	Surprisal: 12.50	Surprisal: 3.38			
RSR^*_{token}: 1.24	RSR^*_{token}: 0.95	RSR^*_{token}: 12.05	RSR^*_{token}: >100	RSR^*_{token}: 0.38	RSR^*_{token}: 8.00	RSR^*_{token}: 1.78			

Table 12: A case study with visualization of token-level rank, surprisal, and RSR values (with rank clipping) across trajectories generated by GPT-OSS-20B, Nemotron-Super-49B-v1.5, and QwQ-32B, as measured by the Qwen-2.5-7B student. Each token is color-coded according to its RSR value, where darker red indicates lower values (preferred under RSR). GPT-OSS-20B exhibits uncommon phrasing and high rank values; Nemotron-Super exhibits conventional reasoning flow and low surprisal; QwQ-32B strikes a more favorable balance between informativeness and alignment, yielding lower token-level RSR values and thus being more likely to be selected. Trajectory-level RSR is defined as the surprisal-weighted average of token-level RSR values with rank clipping (§ 3.4).

and depth. Their tokens show relatively lower rank values alongside higher surprisals, striking a better balance between alignment and informativeness. Consequently, they yield lower token-level RSR values and are preferred under the RSR criterion. These observations are consistent with the distillation results (Table 1) and the subsequent analysis presented in the main body of the paper.

We also observe that token-level RSR may become numerically unstable when extremely low surprisal yields inflated RSR values (e.g., >100). This motivates our surprisal-weighted averaging for trajectory-level RSR (§ 3.4), which stabilizes the metric and improves its effectiveness.

Additional trajectory examples from 11 teachers are available in our released datasets on Hugging Face.

E More Results and Analysis

E.1 Additional Ablation Study

We conduct additional ablation studies with more design variants of RSR. Results are shown in Table 13.

Variants	Avg. Corr.	Δ
Rank-Surprisal Ratio	0.856	
Fixed student	0.785	−0.071
Simple dataset-level average	0.838	−0.018
Use Rank$^{1.05}$	0.845	−0.011
Use Rank$^{0.95}$	0.787	−0.069
Use Surprisal$^{1.05}$	0.847	−0.009
Use Surprisal$^{0.95}$	0.873	0.017
Avg-Rank (clipped)	0.552	−0.304
Avg-Entropy	0.495	−0.361
Surprisal-Weighted Sum of Rank	0.604	−0.252
Rank-Weighted Sum of Surprisal	0.584	−0.272
Rank Minus Surprisal	0.585	−0.271
Rank Times Surprisal	0.538	−0.318
Rank-Entropy Ratio	0.764	−0.092

Table 13: Additional ablation study for *Rank-Surprisal Ratio*. Δ denotes the change in average correlation. All ranks are clipped at $r_{max} = 100$.

The "Fixed student" variant computes RSR using a fixed model (Qwen-3-14B) instead of the target student, and the resulting drop in correlation highlights the importance of student-specific estimation. The "Simple dataset-level average" variant computes the dataset-level RSR by simply averaging trajectory-level scores, as noted in § A.8. The slight degradation in correlation indicates that a simple average of trajectory-level RSR remains a robust estimator of dataset-level suitability, while the surprisal-weighted averaging scheme improves the reliability of the aggregated metric at both trajectory-level and dataset-level.

The "Avg-Rank (clipped)" variant applies rank clipping with $r_{max} = 100$ to Avg-Rank (see Table 4); however, it still exhibits a notable performance gap compared with RSR. This suggests that rank information alone is insufficient to yield strong correlation, underscoring the importance of jointly leveraging rank and surprisal. Similarly, other metrics, such as average entropy and weighted-sum combinations, yield substantially lower correlation than RSR. Our analysis suggests that these alternatives still tend to emphasize high-likelihood trajectories while insufficiently capturing informativeness.

The "Rank Minus Surprisal" metric is computed by subtracting surprisal from the (clipped) rank. Although it models the relationship between rank and surprisal, this simple subtraction does not align well with post-training performance; similarly, "Rank Times Surprisal" also shows poor alignment. The "Rank-Entropy Ratio" is defined as the ratio between the average (clipped) token rank and the average token entropy, where entropy is loosely related to surprisal. Notably, it exhibits a smaller performance drop, suggesting that ratio-based formulations better balance alignment and informativeness.

We also experiment with different exponent choices when computing *Rank-Surprisal Ratio* (with the default setting corresponding to a power of 1 for both rank and surprisal), for example "use Rank$^{1.05}$". The results indicate that our metric is robust to these variations and may yield higher correlation with hyperparameter tuning. Nevertheless, we use the default power-1 setting to keep the formulation simple.

E.2 Additional Results for Correlation Analysis

Table 14 reports both Spearman and Pearson correlations for *Rank-Surprisal Ratio*. The results show that RSR also exhibits strong Pearson correlation with post-training performance. The rationale for primarily using Spearman correlation rather than Pearson correlation is discussed in § A.7.

Metric	Correlation Measure	Student Models					
		Qwen-3-14B	LLaMA-3.1-8B	Qwen-2.5-7B	Qwen-3-4B	Qwen-2.5-3B	**Average**
Rank-Surprisal Ratio	Spearman	0.855	0.845	0.918	0.818	0.845	0.856
	Pearson	0.654	0.880	0.805	0.819	0.811	0.794

Table 14: Comparison of Spearman and Pearson correlation results on *Rank-Surprisal Ratio* (absolute values).

Variant Selection Methods	Qwen-3-14B	LLaMA-3.1-8B	Qwen-2.5-7B	Qwen-3-4B	Qwen-2.5-3B	Average
	Math Avg.	Math Avg.	Math Avg.	Math Avg.	Math Avg.	
Rank-Surprisal Ratio$_{min}$ (5k)	**78.6**	**28.5**	**53.2**	**61.4**	34.8	**51.3**
With Correctness Filtering	77.5	27.9	52.3	60.7	**34.9**	50.7
Fewer Candidates per Teacher	77.6	27.8	52.6	60.9	33.8	50.5
No Selection (55k combined)	72.7	41.5	54.2	58.9	35.3	52.5

Table 15: Performance of RSR-based trajectory selection variants and a No-Selection baseline trained on a 55k combined trajectory dataset. "Math Avg." denotes the average performance over AIME'24, AIME'25, AMC'23, and MATH500.

E.3 Additional Results for Trajectory Selection

E.3.1 RSR-based Trajectory Selection Variants and No-Selection Baseline

Although trajectory selection enables data-efficient training and offers substantial practical value under resource-constrained settings, we are also interested in how it compares with training on all available data without any selection when resources are sufficient. To this end, we construct a "No Selection" baseline. Unlike RSR, which selects 5k trajectories from different teachers, this baseline directly merges 11 trajectory datasets—one per teacher—into a unified training set comprising 55k samples. We do not use the full 165k dataset (i.e., all three trajectory datasets per teacher) for two reasons. First, the computational cost would be prohibitively high. Second, training on multiple highly similar trajectory datasets from the same teacher may reduce data diversity and increase the risk of overfitting. We consider the 55k setting to be sufficiently representative and more consistent with practical use cases.

From Table 15, we observe that the 5k trajectories selected by RSR achieve post-training performance that is comparable to—or even exceeds—the 55k-sample "No Selection" baseline on most student models. An exception is LLaMA-3.1-8B, which may require larger-scale data to develop reasoning ability due to its relatively limited prior exposure to reasoning tasks. In contrast, for stronger students (Qwen-3-14B and Qwen-3-4B), the selected 5k trajectories significantly outperform the unfiltered 55k dataset. This suggests that a small amount of high-quality data is sufficient to effectively activate the reasoning capabilities of student models. Overall, these findings demonstrate that trajectory selection not only improves data efficiency but can also deliver superior performance with substantially fewer resources, further highlighting the practical value of RSR.

Table 15 also presents trajectory selection results for two selection method variants based on RSR, offering further insights.

For "With Correctness Filtering", we consider a combined setting that uses both RSR and verified correctness for trajectory selection. Specifically, for problems with verifiably correct trajectories, we first discard incorrect ones and then select among the remaining correct trajectories based on RSR. The results show no significant improvement over selecting trajectories solely based on RSR, particularly for larger student models. This suggests that correctness is sometimes less critical than overall data suitability.

For "Fewer Candidates per Teacher", we evaluate an 11-to-1 setting in which each candidate pool contains 11 teacher trajectories (one per teacher), instead of the original 33-to-1 setting (three per teacher). This setting focuses on selecting the best trajectory for each problem across different teachers, rather than across multiple generations from the same teacher. The results are comparable, with a slight performance gap relative to the 33-to-1 setting, indicating that RSR remains effective even when each teacher provides only a single trajectory. The results also suggest that while RSR captures

Selection Methods	Qwen-3-14B	LLaMA-3.1-8B	Qwen-2.5-7B	Qwen-3-4B	Qwen-2.5-3B	Average
	GPQA	GPQA	GPQA	GPQA	GPQA	
Random	48.5	26.8	35.4	43.4	22.2	35.3
LLM-judged Quality$_{max}$	53.5	27.3	35.4	**45.5**	21.2	36.6
Rank-Surprisal Ratio$_{min}$	**55.1**	**31.3**	**38.9**	**45.5**	**31.3**	**40.4**

Table 16: Comparison of different trajectory selection methods on the GPQA-Diamond benchmark across student models. We report Acc@4 as the evaluation metric.

Selected Datasets	Data Composition over Teacher Models											Metrics	
	R1	Q3-235B	GPT-120B	Nemotron	QwQ	Q3-30B	Magistral	GPT-20B	Phi-4	Q3-8B	Q3-4B	RSR	Length
RSR$_{min}$ on Q3-14B	6.2%	4.5%	0.1%	0.1%	67.3%	6.7%	1.4%	0.0%	2.9%	2.1%	8.6%	2.57	9363
RSR$_{min}$ on L3.1-8B	5.5%	1.0%	0.0%	3.7%	48.8%	3.9%	9.1%	0.1%	0.2%	21.3%	6.5%	2.69	9939
RSR$_{min}$ on Q2.5-7B	4.1%	1.0%	0.7%	1.7%	55.7%	5.3%	5.9%	0.2%	0.8%	17.6%	6.9%	2.67	9845
RSR$_{min}$ on Q3-4B	2.7%	2.2%	0.0%	0.3%	76.6%	4.1%	3.1%	0.0%	0.6%	4.5%	6.1%	2.56	9419
RSR$_{min}$ on Q2.5-3B	2.4%	0.8%	0.0%	2.4%	45.1%	4.0%	12.2%	0.0%	0.2%	26.0%	6.6%	2.73	10169

Table 17: Data composition and metric statistics of trajectory datasets selected by RSR in the trajectory selection experiments (§ 5.1) for different student models. Model names are abbreviated as Q for Qwen and L for LLaMA.

suitability differences across generations from the same teacher, such differences are less pronounced than those across different teacher models.

E.3.2 Additional Evaluation on GPQA

To more comprehensively evaluate the impact of different trajectory selection methods on post-trained models' reasoning capabilities beyond mathematical problems, we conduct additional evaluation on the GPQA-Diamond benchmark (Rein et al., 2024). GPQA-Diamond consists of 198 challenging multiple-choice questions spanning biology, physics, and chemistry.

Table 16 summarizes the evaluation results of different trajectory selection methods on GPQA-Diamond. Although the results are less stable than those on mathematical benchmarks due to the out-of-domain nature of this evaluation, datasets selected by RSR still achieve the best overall post-training performance. This suggests that RSR can identify suitable trajectories that consistently improve student models' general reasoning capabilities, even when training solely on mathematical problems.

Complete tables combining mathematical and GPQA evaluation results can be found in § F.

E.3.3 Analysis of Datasets Selected by RSR

Table 17 shows the data composition of datasets selected by RSR across 11 teacher models. The resulting distributions vary across student models, demonstrating the metric's ability to select different teacher trajectories tailored to different students. QwQ-32B is generally preferred across student models, consistent with its stable performance. For a clearer contrast in data composition among student models, we refer readers to § E.3.4, where trajectory selection is performed with fewer teachers and consistently strong teachers such as QwQ-32B are removed. We release the RSR-selected datasets for five student models on Hugging Face.

The average RSR values of the selected datasets are also reported in Table 17. These datasets consistently achieve substantially lower RSR values than the teacher trajectory datasets (see § F), validating that our selection procedure effectively identifies trajectories with low RSR for each problem.

E.3.4 Additional Trajectory Selection Experiments with Fewer Teacher Models

Selection Methods	Qwen-3-14B	Qwen-2.5-7B
	Math Avg.	Math Avg.
Random	72.8	47.3
LLM-judged Quality$_{max}$	74.4	48.3
Rank-Surprisal Ratio$_{min}$	**76.8**	**50.0**

Table 18: Comparison of different trajectory selection methods under a reduced-teacher setting. "Math Avg." denotes the average over AIME'24, AIME'25, AMC'23, and MATH500.

To better reflect practical scenarios in which only a few teachers' trajectories are available and generally suitable teachers may be absent, we conduct additional experiments that select trajectories from a reduced set of teachers. Specifically, we select trajectories from candidates gen-

Selected Datasets	Data Composition over Teacher Models						
	Deepseek-R1	Qwen-3-235B	Nemotron	Qwen-3-30B	Magistral	GPT-OSS-20B	Qwen-3-8B
RSR_{min} on Qwen-3-14B	27.40%	21.00%	1.42%	**28.84%**	4.92%	0.24%	16.18%
RSR_{min} on Qwen-2.5-7B	14.52%	8.60%	6.94%	20.76%	10.64%	0.14%	**38.40%**

Table 19: Data composition of trajectory datasets selected by RSR under a reduced-teacher setting (see the setting in § E.3.4 and results in Table 18).

erated by seven teachers: DeepSeek-R1, Qwen-3-235B-Thinking, Nemotron-Super, Qwen-3-30B-Thinking, Magistral-Small, GPT-OSS-20B, and Qwen-3-8B. This teacher set is formed by combining the representative teachers used in Table 2 and § 5.2.

The results are shown in Table 18. Datasets selected by RSR still achieve superior post-training reasoning performance compared with the baselines, demonstrating the effectiveness of our metric when only a limited number of teachers are available. We also observe that the performance gap narrows when selecting from seven teachers compared with eleven teachers, which is expected and suggests that a larger candidate space enables trajectory selection to more effectively identify high-quality training data.

Moreover, Table 19 presents the data composition of datasets selected by RSR from the reduced set of seven teacher models. The distributions differ markedly for Qwen-3-14B and Qwen-2.5-7B: the former student model tends to favor teachers such as DeepSeek-R1 and Qwen-3-30B, whereas the latter student model shows a stronger preference for smaller models such as Qwen-3-8B. These results further demonstrate the effectiveness of RSR in selecting teacher trajectories that are well suited to specific student models.

F Complete Results Tables

Table 25, 26, 27, 28, and 29 present the full metric assessment results across different teacher trajectory datasets on Qwen-3-14B, LLaMA-3.1-8B, Qwen-2.5-7B, Qwen-3-4B, and Qwen-2.5-3B, respectively.

The complete trajectory selection evaluation results underlying Table 6 are presented in Table 20, 21, 22, 23, and 24.

G Others

G.1 License for Artifacts and Data Consent

All artifacts used in this paper are publicly available for academic research purposes, including AIME, AMC, MATH500, GPQA-Diamond and Numina-Math.

G.2 Data Statement

The training datasets consist solely of mathematics problems and solutions and contain no offensive content or personal information.

G.3 AI Assistant Usage Statement

We used ChatGPT for writing refinement and minor coding assistance. AI assistants were not involved in research innovation, and all core contributions were developed solely by the authors.

Selection Methods	Qwen-3-14B					
	AIME'24	AIME'25	AMC'23	MATH500	**Math Avg.**	GPQA-Diamond
Random	59.2	46.7	86.3	88.6	70.2	48.5
Token Length$_{max}$	61.7	51.7	87.5	84.8	71.4	–
Rule-based Quality$_{max}$	58.3	47.5	91.3	92.0	72.3	–
LLM-judged Quality$_{max}$	60.0	49.2	90.6	93.6	73.4	53.5
Surprisal$_{min}$	62.5	50.0	88.1	92.8	73.4	–
G-Norm$_{min}$	59.2	50.0	89.4	92.4	72.7	–
Rank-Surprisal Ratio$_{min}$	67.5	59.2	93.1	94.6	**78.6**	**55.1**

Table 20: Full post-training reasoning evaluation results for trajectory selection on Qwen-3-14B. "Math Avg." denotes the average over AIME'24, AIME'25, AMC'23, and MATH500.

Selection Methods	LLaMA-3.1-8B					
	AIME'24	AIME'25	AMC'23	MATH500	**Math Avg.**	GPQA-Diamond
Random	2.5	5.8	29.4	50.8	22.1	26.8
Token Length$_{max}$	8.3	6.7	36.9	57.4	27.3	–
Rule-based Quality$_{max}$	6.7	9.2	29.4	58.0	25.8	–
LLM-judged Quality$_{max}$	1.7	4.2	38.1	58.6	25.6	27.3
Surprisal$_{min}$	5.0	6.7	25.6	56.8	23.5	–
G-Norm$_{min}$	5.8	4.2	36.9	57.6	26.1	–
Rank-Surprisal Ratio$_{min}$	5.0	8.3	36.9	63.6	**28.5**	**31.3**

Table 21: Full post-training reasoning evaluation results for trajectory selection on LLaMA-3.1-8B. "Math Avg." denotes the average over AIME'24, AIME'25, AMC'23, and MATH500.

Selection Methods	Qwen-2.5-7B					
	AIME'24	AIME'25	AMC'23	MATH500	**Math Avg.**	GPQA-Diamond
Random	18.3	19.2	62.5	82.8	45.7	35.4
Token Length$_{max}$	22.5	21.7	61.9	75.6	45.4	–
Rule-based Quality$_{max}$	24.2	25.0	72.5	84.8	51.6	–
LLM-judged Quality$_{max}$	30.0	23.3	66.9	87.0	51.8	35.4
Surprisal$_{min}$	22.5	20.8	63.1	79.2	46.4	–
G-Norm$_{min}$	27.5	25.0	66.3	79.2	49.5	–
Rank-Surprisal Ratio$_{min}$	29.2	25.8	71.3	86.6	**53.2**	**38.9**

Table 22: Full post-training reasoning evaluation results for trajectory selection on Qwen-2.5-7B. "Math Avg." denotes the average over AIME'24, AIME'25, AMC'23, and MATH500.

Selection Methods	Qwen-3-4B					
	AIME'24	AIME'25	AMC'23	MATH500	**Math Avg.**	GPQA-Diamond
Random	30.8	30.8	68.8	85.2	53.9	43.4
Token Length$_{max}$	28.3	28.3	71.9	76.6	51.3	–
Rule-based Quality$_{max}$	36.7	32.5	77.5	85.4	58.0	–
LLM-judged Quality$_{max}$	37.5	32.5	77.5	88.8	59.1	**45.5**
Surprisal$_{min}$	29.2	33.3	71.9	78.8	53.3	–
G-Norm$_{min}$	34.2	33.3	79.4	89.4	59.1	–
Rank-Surprisal Ratio$_{min}$	44.2	35.0	77.5	88.8	**61.4**	**45.5**

Table 23: Full post-training reasoning evaluation results for trajectory selection on Qwen-3-4B. "Math Avg." denotes the average over AIME'24, AIME'25, AMC'23, and MATH500.

Selection Methods	Qwen-2.5-3B					
	AIME'24	AIME'25	AMC'23	MATH500	**Math Avg.**	GPQA-Diamond
Random	6.7	7.5	36.3	61.2	27.9	22.2
Token Length$_{max}$	5.0	10.0	37.5	56.0	27.1	–
Rule-based Quality$_{max}$	5.8	10.0	45.0	63.8	31.2	–
LLM-judged Quality$_{max}$	12.5	10.8	39.4	68.4	32.8	21.2
Surprisal$_{min}$	5.8	8.3	39.4	62.0	28.9	–
G-Norm$_{min}$	9.2	9.2	46.3	59.0	30.9	–
Rank-Surprisal Ratio$_{min}$	11.7	11.7	45.0	70.8	**34.8**	**31.3**

Table 24: Full post-training reasoning evaluation results for trajectory selection on Qwen-2.5-3B. "Math Avg." denotes the average over AIME'24, AIME'25, AMC'23, and MATH500.

Metrics	Deepseek-R1	Q3-235B	GPT-120B	Nemotron	QwQ-32B	Q3-30B	Magistral	GPT-20B	Phi-4	Q3-8B	Q3-4B
Avg-Token Length	12077.7	12571.1	2552.3	8798.2	9070.4	10887.1	10993.3	3822.7	3643.1	10473.9	13145.8
Teacher Performance	0.911	0.912	0.883	0.823	0.852	0.923	0.710	0.834	0.727	0.825	0.873
Verified Accuracy	0.849	0.859	0.801	0.786	0.820	0.857	0.776	0.784	0.804	0.803	0.835
Rule-based Quality	0.226	-0.031	-0.484	0.259	0.395	-0.106	0.100	-0.393	-0.389	0.465	-0.042
LLM-judged Quality	0.908	0.966	0.896	0.882	0.901	0.963	0.815	0.863	0.823	0.911	0.951
G-Norm	33.615	33.581	72.106	31.801	39.701	34.569	35.455	66.665	62.448	33.913	33.580
Influence Score ($\times 10^5$)	0.161	0.696	0.418	1.447	0.795	0.671	-0.008	0.244	-0.348	0.816	0.633
Avg-Surprisal	0.660	0.616	1.162	0.424	0.629	0.591	0.410	1.276	1.068	0.485	0.581
Avg-Rank	49.412	55.609	430.379	65.293	55.249	58.448	41.968	365.886	303.840	45.098	49.145
Avg-Rank (clipped)	1.930	1.811	4.097	1.422	1.682	1.728	1.355	4.652	3.589	1.457	1.695
GRACE	0.028	0.028	0.168	0.025	0.031	0.030	0.025	0.154	0.113	0.023	0.026
Avg-RSR$_{\text{token}}$ ($\times 10^8$)	1.564	2.970	0.706	2.598	0.768	4.078	31.957	0.348	0.315	2.463	2.569
Avg-RSR$^{\text{filter}}_{\text{token}}$	8.713	9.962	62.383	13.663	9.674	10.437	22.701	53.858	43.930	8.799	9.079
RSR (200 sample)	2.916	2.943	3.504	3.342	2.684	2.915	3.252	3.679	3.348	2.961	2.904
RSR	2.925	2.940	3.527	3.352	2.673	2.923	3.302	3.645	3.360	3.003	2.918
Post-Training Performance	77.1	71.8	66.7	72.2	77.4	77.2	68.8	69.5	54.1	74.6	76.8

Table 25: Full metric assessment results on Qwen-3-14B. Model names are abbreviated with Q for Qwen.

Metrics	Deepseek-R1	Q3-235B	GPT-120B	Nemotron	QwQ-32B	Q3-30B	Magistral	GPT-20B	Phi-4	Q3-8B	Q3-4B
Avg-Token Length	12077.7	12571.1	2552.3	8798.2	9070.4	10887.1	10993.3	3822.7	3643.1	10473.9	13145.8
Teacher Performance	0.911	0.912	0.883	0.823	0.852	0.923	0.710	0.834	0.727	0.825	0.873
Verified Accuracy	0.849	0.859	0.801	0.786	0.820	0.857	0.776	0.784	0.804	0.803	0.835
Rule-based Quality	0.226	-0.031	-0.484	0.259	0.395	-0.106	0.100	-0.393	-0.389	0.465	-0.042
LLM-judged Quality	0.908	0.966	0.896	0.882	0.901	0.963	0.815	0.863	0.823	0.911	0.951
G-Norm	52.768	55.501	120.459	55.882	56.478	57.108	43.666	109.119	95.909	48.365	53.058
Influence Score ($\times 10^6$)	2.865	1.330	2.361	1.123	1.359	1.258	0.932	2.755	2.205	1.125	1.243
Avg-Surprisal	0.945	0.927	1.418	0.724	0.953	0.899	0.668	1.530	1.277	0.754	0.866
Avg-Rank	10.328	11.101	72.356	11.762	11.028	11.206	8.463	66.664	49.643	8.721	9.757
Avg-Rank (clipped)	2.831	2.821	5.633	2.183	2.687	2.665	2.016	6.178	4.638	2.174	2.552
GRACE	0.162	0.177	1.005	0.185	0.183	0.181	0.111	0.853	0.639	0.143	0.159
Avg-RSR$_{\text{token}}$ ($\times 10^8$)	1.073	0.872	0.588	0.976	0.578	0.972	1.843	0.284	0.555	0.727	0.934
Avg-RSR$^{\text{filter}}_{\text{token}}$	3.604	3.791	18.711	4.019	3.706	3.805	7.881	17.525	13.173	3.206	3.458
RSR (200 sample)	2.995	3.044	3.976	2.997	2.848	2.960	3.000	4.104	3.608	2.857	2.946
RSR	2.996	3.044	3.971	3.016	2.818	2.965	3.020	4.038	3.633	2.882	2.945
Post-Training Performance	28.1	22.0	15.2	23.7	27.1	26.7	22.8	17.9	14.5	26.5	28.2

Table 26: Full metric assessment results on LLaMA-3.1-8B. Model names are abbreviated with Q for Qwen.

Metrics	Deepseek-R1	Q3-235B	GPT-120B	Nemotron	QwQ-32B	Q3-30B	Magistral	GPT-20B	Phi-4	Q3-8B	Q3-4B
Avg-Token Length	12077.7	12571.1	2552.3	8798.2	9070.4	10887.1	10993.3	3822.7	3643.1	10473.9	13145.8
Teacher Performance	0.911	0.912	0.883	0.823	0.852	0.923	0.710	0.834	0.727	0.825	0.873
Verified Accuracy	0.849	0.859	0.801	0.786	0.820	0.857	0.776	0.784	0.804	0.803	0.835
Rule-based Quality	0.226	-0.031	-0.484	0.259	0.395	-0.106	0.100	-0.393	-0.389	0.465	-0.042
LLM-judged Quality	0.908	0.966	0.896	0.882	0.901	0.963	0.815	0.863	0.823	0.911	0.951
G-Norm	42.327	39.127	84.024	38.592	45.587	40.408	32.527	78.025	73.562	38.400	37.860
Influence Score ($\times 10^6$)	-0.520	-0.766	-0.812	-0.732	-0.737	-0.767	-0.466	-0.788	-0.740	-0.727	-0.770
Avg-Surprisal	0.825	0.799	1.236	0.597	0.820	0.767	0.553	1.356	1.131	0.647	0.748
Avg-Rank	6.192	6.438	36.628	6.413	6.312	6.411	4.724	35.233	25.818	5.000	5.683
Avg-Rank (clipped)	2.477	2.416	4.557	1.842	2.280	2.264	1.710	5.189	3.921	1.869	2.198
GRACE	0.168	0.199	1.760	0.126	0.160	0.142	0.119	1.491	0.908	0.120	0.139
Avg-$\text{RSR}_{\text{token}}$ ($\times 10^7$)	2.980	2.694	1.104	5.508	3.055	2.253	3.582	0.515	0.824	2.009	2.590
Avg-$\text{RSR}^{\text{filter}}_{\text{token}}$	2.671	2.709	9.663	3.063	2.542	2.705	4.359	9.553	7.163	2.434	2.551
RSR (200 sample)	2.999	3.030	3.670	3.066	2.803	2.950	3.067	3.887	3.471	2.864	2.935
RSR	3.002	3.023	3.686	3.086	2.779	2.951	3.091	3.827	3.468	2.888	2.940
Post-Training Performance	47.3	45.0	40.7	48.3	52.0	50.0	47.6	42.7	35.2	52.0	51.8

Table 27: Full metric assessment results on Qwen-2.5-7B. Model names are abbreviated with Q for Qwen.

Metrics	Deepseek-R1	Q3-235B	GPT-120B	Nemotron	QwQ-32B	Q3-30B	Magistral	GPT-20B	Phi-4	Q3-8B	Q3-4B
Avg-Token Length	12077.7	12571.1	2552.3	8798.2	9070.4	10887.1	10993.3	3822.7	3643.1	10473.9	13145.8
Teacher Performance	0.911	0.912	0.883	0.823	0.852	0.923	0.710	0.834	0.727	0.825	0.873
Verified Accuracy	0.849	0.859	0.801	0.786	0.820	0.857	0.776	0.784	0.804	0.803	0.835
Rule-based Quality	0.226	-0.031	-0.484	0.259	0.395	-0.106	0.100	-0.393	-0.389	0.465	-0.042
LLM-judged Quality	0.908	0.966	0.896	0.882	0.901	0.963	0.815	0.863	0.823	0.911	0.951
G-Norm	33.263	33.036	70.922	35.749	43.248	33.884	39.333	67.150	62.701	38.211	33.109
Influence Score ($\times 10^4$)	-0.284	-0.013	-1.077	3.601	1.919	0.068	-1.214	-1.202	-0.965	1.395	-1.149
Avg-Surprisal	0.721	0.685	1.208	0.474	0.693	0.650	0.450	1.319	1.102	0.526	0.615
Avg-Rank	9.172	10.267	82.658	10.747	9.773	10.346	7.503	72.151	54.562	7.710	9.011
Avg-Rank (clipped)	2.132	2.023	4.581	1.526	1.835	1.896	1.422	5.130	3.955	1.535	1.802
GRACE	0.148	0.151	0.524	0.148	0.165	0.145	0.139	0.436	0.425	0.124	0.116
Avg-$\text{RSR}_{\text{token}}$ ($\times 10^8$)	1.314	1.785	0.361	4.915	1.122	2.187	29.884	0.203	0.182	3.123	2.229
Avg-$\text{RSR}^{\text{filter}}_{\text{token}}$	3.641	4.058	21.345	6.773	3.619	4.197	16.064	18.829	14.369	4.010	3.912
RSR (200 sample)	2.947	2.961	3.771	3.197	2.660	2.908	3.124	3.937	3.579	2.881	2.919
RSR	2.958	2.955	3.794	3.216	2.649	2.917	3.160	3.888	3.588	2.919	2.928
Post-Training Performance	55.8	53.4	47.9	56.4	61.2	58.8	52.2	48.4	40.2	61.2	61.9

Table 28: Full metric assessment results on Qwen-3-4B. Model names are abbreviated with Q for Qwen.

Metrics	Deepseek-R1	Q3-235B	GPT-120B	Nemotron	QwQ-32B	Q3-30B	Magistral	GPT-20B	Phi-4	Q3-8B	Q3-4B
Avg-Token Length	12077.7	12571.1	2552.3	8798.2	9070.4	10887.1	10993.3	3822.7	3643.1	10473.9	13145.8
Teacher Performance	0.911	0.912	0.883	0.823	0.852	0.923	0.710	0.834	0.727	0.825	0.873
Verified Accuracy	0.849	0.859	0.801	0.786	0.820	0.857	0.776	0.784	0.804	0.803	0.835
Rule-based Quality	0.226	-0.031	-0.484	0.259	0.395	-0.106	0.100	-0.393	-0.389	0.465	-0.042
LLM-judged Quality	0.908	0.966	0.896	0.882	0.901	0.963	0.815	0.863	0.823	0.911	0.951
G-Norm	29.027	27.225	74.196	27.733	30.065	27.843	26.490	67.096	62.386	25.996	26.518
Influence Score ($\times 10^5$)	-1.288	-2.927	-2.157	-2.873	-3.711	-2.730	-1.694	-2.311	-2.555	-3.225	-2.812
Avg-Surprisal	0.903	0.885	1.346	0.657	0.891	0.847	0.608	1.454	1.210	0.704	0.822
Avg-Rank	18.247	19.034	145.459	24.048	21.422	19.906	16.483	119.646	90.940	16.967	17.978
Avg-Rank (clipped)	2.794	2.747	5.307	2.040	2.550	2.551	1.855	5.870	4.381	2.049	2.462
GRACE	0.096	0.092	0.602	0.093	0.100	0.147	0.077	0.991	0.426	0.075	0.085
Avg-$\text{RSR}_{\text{token}}$ ($\times 10^8$)	0.953	0.805	0.177	2.848	1.916	0.746	2.188	0.0768	1.814	0.723	0.791
Avg-$\text{RSR}^{\text{filter}}_{\text{token}}$	4.829	4.977	29.896	6.096	5.273	5.092	8.041	24.997	18.963	4.517	4.751
RSR (200 sample)	3.094	3.109	3.929	3.084	2.885	3.017	3.029	4.095	3.603	2.891	2.991
RSR	3.095	3.103	3.944	3.107	2.860	3.012	3.050	4.037	3.622	2.911	2.994
Post-Training Performance	29.6	26.4	22.9	33.0	33.0	31.2	30.6	24.4	18.2	34.2	33.3

Table 29: Full metric assessment results on Qwen-2.5-3B. Model names are abbreviated with Q for Qwen.

Evaluation prompt for LLM-judged quality assessment

```
You are a meticulous and highly critical evaluator of AI reasoning. Your primary goal is to
    identify and quantify subtle flaws, logical gaps, inefficiencies, and hidden assumptions.
    Do not default to a high score. Your starting assumption should be critical, and you must
    rigorously justify every point awarded.

First, please carefully read the following problem statement:
<Problem>
{question}
</Problem>

Now, please carefully read the following candidate's chain-of-thought reasoning:
<Reasoning>
{reasoning_to_evaluate}
</Reasoning>

When evaluating this reasoning, you must adhere to the following five key evaluation criteria
    and the scoring rubric below.

Scoring Guidelines and Calibration:
You must use the full 0.0 to 1.0 scale. Scores should not be clustered at the top. Use this
    rubric to anchor your scores:
1.0 (Exceptional/Flawless): Reserved for reasoning that is not only correct but also elegant,
    insightful, and comprehensive. It is perfectly structured and leaves no room for doubt.
    This score should be exceedingly rare.
0.8 - 0.9 (Excellent but Imperfect): The core reasoning is valid and well-supported, but there
    may be very minor, superficial issues (e.g., a trivial typo in a formula that doesn't
    affect the outcome, a slightly awkward phrasing). The conclusion is unaffected.
0.5 - 0.7 (Competent but Flawed): The reasoning is generally on the right track but contains
    noticeable and non-trivial flaws. Examples include: a minor factual error, a logical leap
    that requires the reader to fill in the blanks, an inefficient method where a much simpler
     one exists, or a partially incomplete answer.
0.2 - 0.4 (Poor): The reasoning contains fundamental flaws that largely invalidate the process
    or conclusion. Examples include: a significant factual error, a clear logical fallacy,
    misunderstanding of the core problem constraints.
0.0 - 0.1 (Unacceptable): The reasoning is completely incorrect, irrelevant, nonsensical, or
    makes no meaningful attempt to solve the problem.

Crucial Instruction for High Scores:
To combat score inflation, you must justify high scores with the same rigor as low scores. For
    any criterion where you assign a score of 0.9 or 1.0, your justification must explicitly
    state what makes the reasoning exceptional and why it lacks even subtle flaws.

Evaluation Criteria:
Factual Accuracy:
Scrutinize every claim, formula, and piece of domain knowledge. Is it precisely correct? Assess
     the application of problem constraints, paying close attention to edge cases and boundary
     conditions. Penalize any inaccuracy, no matter how small.

Logical Rigor:
Probe for hidden assumptions and unstated premises. Does each conclusion necessarily and
    unambiguously follow from the preceding steps? Identify any logical fallacies,
    contradictions, or jumps in reasoning. A chain is only as strong as its weakest link.

Solution Completeness:
Does the reasoning address all parts of the problem statement exhaustively? Does it consider
    all possible cases, sub-problems, and nuances? An answer that is correct for one case but
    ignores others is incomplete.

Reasoning Efficiency:
Is this the most direct and economical path to the solution? Penalize any unnecessary
    complexity, redundant steps, or exploration of irrelevant tangents, even if they
    eventually lead to the correct answer. The cognitive effort should be proportionate to the
     problem's complexity.

Presentation Quality:
How clearly is the reasoning communicated? Is the structure logical and easy to follow?
    Ambiguous language, poor organization, or a confusing sequence of steps should be
```

```
    penalized. An observer should be able to verify the reasoning process without difficulty.

For each of the five evaluation criteria, please give a score from 0.0 to 1.0 (in 0.1
    increments) and a brief, clear justification for that score in the JSON structure.

Your output must be a single, valid JSON object. The format of the JSON object is as follows:
```json
{
  "dimensional_evaluation": {
    "factual_accuracy": {
      "score": <float between 0.0 and 1.0>,
      "reason": "<Your justification for the factual accuracy score>"
    },
    "logical_rigor": {
      "score": <float between 0.0 and 1.0>,
      "reason": "<Your justification for the logical rigor score>"
    },
    "solution_completeness": {
      "score": <float between 0.0 and 1.0>,
      "reason": "<Your justification for the solution completeness score>"
    },
    "reasoning_efficiency": {
      "score": <float between 0.0 and 1.0>,
      "reason": "<Your justification for the reasoning efficiency score>"
    },
    "presentation_quality": {
      "score": <float between 0.0 and 1.0>,
      "reason": "<Your justification for the presentation quality score>"
    }
  },
  "overall_score": <float between 0.0 and 1.0>,
  "overall_reason": "<A concise summary justifying the overall score by synthesizing the key
    findings from the dimensional evaluation.>"
}
```

Table 30: Evaluation prompt for LLM-judged quality assessment as a baseline metric.
```