On the Step Length Confounding in LLM Reasoning Data Selection

Bing Wang[1,2,3], **Rui Miao**[3,4], **Chen Shen**[3*], **Shaotian Yan**[3], **Kaiyuan Liu**[3,5], **Ximing Li**[1,2,7*], **Xiaosong Yuan**[1,2,3], **Sinan Fan**[3,5], **Jun Zhang**[6], **Jieping Ye**[3]

[1] College of Computer Science and Technology, Jilin University

[2] Key Laboratory of Symbolic Computation and Knowledge Engineering, MoE, Jilin University

[3] Alibaba Cloud Computing [4] School of Artificial Intelligence, Jilin University

[5] College of Computer Science and Technology, Zhejiang University

[6] Department of Mathematics, University of Michigan [7] RIKEN Center for Advanced Intelligence Project

{wangbing1416,zjushenchen,liximing86}@gmail.com, miaorui24@mails.jlu.edu.cn

Abstract

Large reasoning models have recently demonstrated strong performance on complex tasks that require long chain-of-thought reasoning, through supervised fine-tuning on large-scale and high-quality datasets. To construct such datasets, existing pipelines generate long reasoning data from more capable Large Language Models (LLMs) and apply manually heuristic or naturalness-based selection methods to filter high-quality samples. Despite the proven effectiveness of *naturalness-based data selection*, which ranks data by the average log probability assigned by LLMs, our analysis shows that, when applied to LLM reasoning datasets, it systematically prefers samples with longer reasoning steps (*i.e.,* more tokens per step) rather than higher-quality ones, a phenomenon we term **step length confounding**. Through quantitative analysis, we attribute this phenomenon to low-probability first tokens in reasoning steps; longer steps dilute their influence, thereby inflating the average log probabilities. To address this issue, we propose two variant methods: ASLEC-DROP, which drops first-token probabilities when computing average log probability, and ASLEC-CASL, which applies a causal debiasing regression to remove the first tokens' confounding effect. Experiments across four LLMs and five evaluation benchmarks demonstrate the effectiveness of our approach in mitigating the step length confounding problem.

1 Introduction

Recently, a variety of large reasoning models, *e.g.,* DeepSeek-R1 (Guo et al., 2025), have achieved remarkable performance on complex reasoning tasks that demand long Chain-of-Thought (CoT) capabilities (Yang et al., 2025a; Team, 2025). To elicit such long CoT reasoning abilities in foundation models, Supervised Fine-Tuning (SFT) on large-scale, high-quality datasets has become a standard paradigm (Chen et al., 2025b; Guha et al., 2025; Zhao et al., 2025; Yuan et al., 2026). Existing approaches typically begin by collecting complex mathematical and scientific problems, and then prompting stronger Large Language Models (LLMs) to generate answers as SFT datasets (Guha et al., 2025; Yuan et al., 2025; Huang et al., 2025). Despite this pipeline effectively scaling up SFT data, such datasets still contain noisy instances, *e.g.,* incorrect reasoning steps (Zheng et al., 2025) or overly complex reasoning trajectories (Sui et al., 2025). To address this issue and build higher-quality data subsets, LLM reasoning data selection has emerged as an active research topic (Ye et al., 2025; Muennighoff et al., 2025).

Generally, existing reasoning data selection methods often rely on heuristic rules, *e.g.,* verifiable answer correctness (Zhao et al., 2025; Wu et al., 2025), response diversity (Jung et al., 2025; Li et al., 2025a), and problem difficulty (Muennighoff et al., 2025; Guha et al., 2025). These methods often depend heavily on manually crafted heuristics and do not consider the trained LLM's adaptability to the SFT data. To overcome this limitation, the community introduces a **naturalness-based data selection** strategy (Zhang et al., 2025; Just et al., 2025), which involves computing the log probability assigned by an LLM to each SFT data sample and selecting those with higher average probabilities, as they are presumed to be better aligned with the LLM's inherent preferences.

Unfortunately, our findings reveal that, when applied to long CoT datasets, *the naturalness-based selection methods significantly prefer samples with longer reasoning steps* (*i.e.,* more tokens per step) rather than higher-adaptability ones. We refer to this phenomenon as the **step length confounding** problem in this work. We show in Fig. 1, the step-length distribution of the *selected* SFT data differs markedly from that of the *unselected* data. To further investigate the cause of this confounder,

*Corresponding authors

Findings of the Association for Computational Linguistics: ACL 2026, pages 18443–18457

we build upon the quantitative results presented in Figs. 2 and 3. We observe that longer reasoning steps generally yield higher average log probabilities. This phenomenon can be explained by prior work showing that *the first token of each reasoning step* often folks into different reasoning branches, thereby exhibiting higher entropy and consequently *lower log probabilities* (Wang et al., 2025; Cheng et al., 2025). *Longer steps, however, dilute the impact of these low-probability first tokens, leading to a higher overall step log probability*, which in turn makes such longer-step examples more likely to be selected.

Given the above conclusion that low-probability first tokens lead to the step length confounding problem, we propose a mitigation method, namely **A**lleviating **S**tep **Le**ngth **C**onfounding (**ASLEC**), which includes two variant approaches **ASLEC-DROP** and **ASLEC-CASL**. Specifically, ASLEC-DROP attempts to mitigate the confounding problem by simply dropping the first-token probabilities when computing the global average log probability. Despite this straightforward approach offering a preliminary mitigation to the confounding issue, it also entirely discards the contribution of the first token to data selection. Accordingly, ASLEC-CASL, inspired by causal debiasing techniques (Udomcharoenchaikit et al., 2022), fits a linear regression model to disentangle the first-token ratio as a confounding factor, and removes its effect when computing the global average log probability.

In our experiments, we train four LLMs of varying sizes on two LLM reasoning benchmark datasets *LIMO-v2* (Ye et al., 2025) and *AceReason-1.1-SFT* (Chen et al., 2025b), and evaluate the performance of different data selection strategies across five evaluation benchmarks. Our results demonstrate that the two proposed variants, ASLEC-DROP and ASLEC-CASL, consistently outperform the state-of-the-art naturalness-based selection method, Local LP (Just et al., 2025), achieving average accuracy improvements of approximately 6.28% and 9.08%, respectively, across all model sizes and datasets. *Our source code and data are released in* https://github.com/wangbing1416/ASLEC. Meanwhile, in our implementation, we sample a large number of multi-source, multi-solution responses for *LIMO-v2* and 10k *AceReason-1.1-SFT* problems (average 64 responses per question). *These large-scale SFT datasets are also be released in* 🤗 https://huggingface.co/collections/wangbing1416/msms-cot-sft.

Generally, our contributions can be summarized as the following three-fold:

- Through extensive experiments, we identify a step length confounding problem in existing naturalness-based LLM reasoning data selection methods, and reveal that the cause lies in the low-probability first token of each step.
- We propose two variant methods, ASLEC-DROP and ASLEC-CASL, which alleviate the step length confounding problem by intervening on the first-token probability when computing the global average log probability.
- Extensive experiments demonstrate the effectiveness of our proposed method and its ability to mitigate step length confounding.

2 Preliminary Experimental Analysis on Step Length Confounding

In this section, our preliminary experiments reveal that existing naturalness-based approaches for LLM reasoning data selection consistently suffer from **step length confounding**: they tend to prefer samples with longer reasoning steps.

2.1 Naturalness-Based Data Selection

Typically, an LLM reasoning SFT dataset is defined as $\mathcal{D} = \{\mathbf{q}_i, \mathbf{c}_i, \mathbf{a}_i\}_{i=1}^{N}$, where $\mathbf{q}$ denotes one question, and $\mathbf{c}$ and $\mathbf{a}$ represent its long CoT reasoning trajectory and final answer, respectively. The SFT objective is to optimize model parameters $\boldsymbol{\theta}$ by minimizing the negative log-likelihood of the target sequence $\mathbf{o}_i =$ <think> $\mathbf{c}_i$ </think> $\mathbf{a}_i$ as:

$$\mathcal{L}_{\text{SFT}}(\boldsymbol{\theta}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{|\mathbf{o}_i|}\log P_{\boldsymbol{\theta}}\left(o_{i,t} \mid o_{i,<t}, \mathbf{q}_i\right),$$

which is equal to the causal LM cross-entropy loss. While SFT typically treats all samples equally, data quality critically influences reasoning performance, as noisy or inconsistent trajectories can mislead learning. This motivates data selection strategies that prefer high-quality and informative subsets $\widehat{\mathcal{D}} \in \mathcal{D}$ to improve robustness. Among these works, **naturalness-based methods** leverage the log probabilities produced by the LLM during SFT to select the data to which the model is best adapted. Formally, three representative methods are as follows:

- **Log probabilities** (Zhang et al., 2025) or **Perplexity** (Muennighoff et al., 2023; Yin and Rush, 2025) computes the geometric mean of the probabilities assigned to the target sequence outputs, as follows:

$$s_i^{\text{logp}} = \frac{1}{|\mathbf{o}_i|} \sum_{t=1}^{|\mathbf{o}_i|} \log P_{\boldsymbol{\theta}} \left(o_{i,t} \mid \mathbf{o}_{i,<t}, \mathbf{q}_i \right),$$
$$s_i^{\text{ppl}} = \exp\left(-s_i^{\text{logp}} \right). \quad (1)$$

A higher s_i^{logp} indicates that the model naturally adapts better to the given data.

- **Local log probabilities** (Just et al., 2025) split the sequence $\mathbf{o}_i$ into steps $\mathcal{S}_i = \{\mathbf{s}_{ij}\}_{j=1}^{|\mathcal{S}_i|}$ by the token \n\n or sentences. For each step, it considers the question and its previous k steps as context and calculates the geometric mean of its log probability accordingly.

$$s_i^{\text{loc}} = \frac{1}{|\mathcal{S}_i|} \sum_{\mathbf{s}_{ij} \in \mathcal{S}_i} \frac{1}{|\mathbf{s}_{ij}|} \sum_{l=1}^{|\mathbf{s}_{ij}|} \log \quad (2)$$
$$P_{\boldsymbol{\theta}} \left(s_{ijl} \mid \mathbf{s}_{ij,<l}, \mathbf{s}_{i,\max(1,j-k):j-1}, \mathbf{q}_i \right).$$

- **Entropy** (Cui et al., 2025; Wang et al., 2025) measures the average token-level uncertainty of the model's predictions.

$$s_i^{\text{etp}} = \frac{1}{|\mathbf{o}_i|} \sum_{t=1}^{|\mathbf{o}_i|} \Big[-\sum\nolimits_{v \in \mathcal{V}} P_{\boldsymbol{\theta}}(v \mid o_{i,<t}, \mathbf{q}_i)$$
$$\log P_{\boldsymbol{\theta}}(v \mid o_{i,<t}, \mathbf{q}_i) \Big], \quad (3)$$

where $\mathcal{V}$ represents the vocabulary, and lower entropy means the model is more confident in its outputs on the given example.

Existing naturalness-based methods typically select a subset $\widehat{\mathcal{D}}$ from the large-scale dataset $\mathcal{D}$ by either highest s_i^{logp} and s_i^{loc}, or lowest s_i^{ppl} and s_i^{etp}.

2.2 Experimental Setup

Models. Our experiments utilize four long CoT reasoning LLMs of different families and parameters, including *QwQ-32B* (Team, 2025), *Qwen3-32B* (Yang et al., 2025a), *DeepSeek-R1-Distill-Qwen-32B* (Guo et al., 2025), and *gpt-oss-120b* (Agarwal et al., 2025), as data sources for generating reasoning SFT data. We then use *Qwen3-4B-Base* (Yang et al., 2025a) as the target LLM to evaluate its log probabilities *w.r.t* the SFT data. Detailed model cards for all LLMs are provided in Appendix B.1.

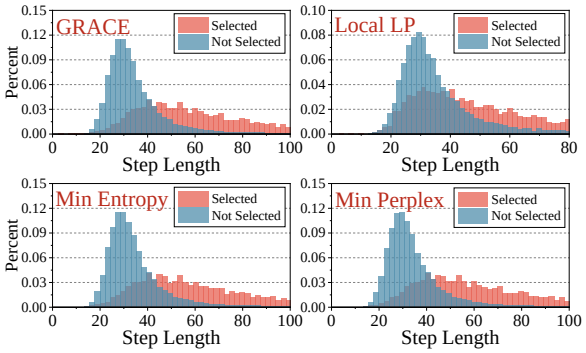

Figure 1: Step length distribution of data samples *selected* and *unselected* by different naturalness-based data selection methods.

Benchmark and evaluation. We conduct our experiments on *LIMO-v2* (Ye et al., 2025), a prevalent LLM reasoning benchmark comprising 800 carefully curated mathematical reasoning problems. For each problem, we generate 5 different responses from each of the 4 LLMs described above, using temperature sampling with $\tau = 0.6$. From the generated $4 \times 5 = 20$ responses per problem, we select 5 final responses using one of the four representative naturalness-based data selection methods, (i) *GRACE* (Zhang et al., 2025), which selects the responses with the highest s^{logp}; (ii) *Local LP* (Just et al., 2025), which also selects the responses with the highest s^{loc}; (iii) *Min Entropy* (Cui et al., 2025), which selects the responses with the smallest s^{etp}; (iv) *Min Perplex*, which select the responses with the smallest s_i^{ppl}, to analyze the step length confounding phenomenon. More details on the benchmarks and experimental setup are provided in Appendix B.2, and additional analysis across more datasets, *e.g., AceReason-1.1-SFT*, and more LLMs can be found in Appendix C.

2.3 Results on Step Length Confounding

Through the preliminary experiments in this section, we find that these naturalness-based methods suffer from the step length confounding issue.

Results and analysis. In Fig. 1, we illustrate the selection difference of naturalness-based methods across the 16,000 responses (800 problems × 20 responses each) generated by the four LLMs. Fig. 1 compares the step-length distributions of responses *selected* versus *not selected* by four naturalness-based data selection methods. Across all methods, selected responses consistently exhibit longer step

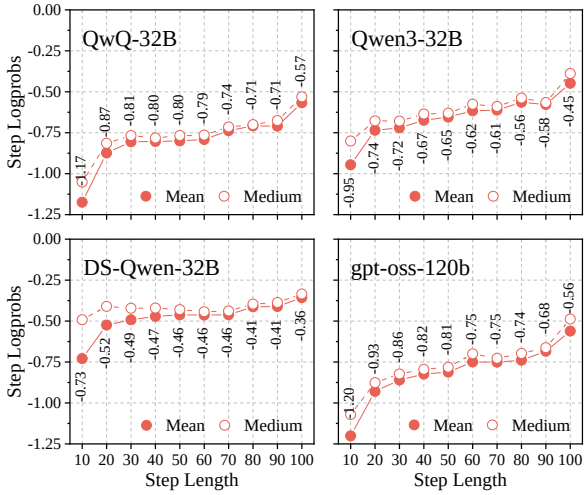

Figure 2: Relationship between step-level log probability and step length.

lengths, whereas the step lengths of unselected responses are more concentrated at lower values, with an average of approximately 30. This pattern underscores the consistent influence of step length on the decisions made by these naturalness-based criteria. [1] Based on these observations, we formulate the following conclusion and refer to this phenomenon as *step length confounding*.

⋆ **Conclusion 1.** The naturalness-based reasoning data selection approach tends to **prefer samples with longer reasoning steps** (*i.e.,* more tokens per step).

2.4 Why Step Length Confounding?

Given the step length confounding problem in LLM reasoning data selection, we seek to figure out the intrinsic causes resulting in this issue. Therefore, we give the following further empirical evidence.

For longer steps, the model assigns higher step-level log probabilities. We first investigate the relationship between step length and the average log probability per step. As illustrated in Fig. 2, outputs from different LLMs are segmented into steps. For steps of different lengths, we compute the average step-level log probabilities assigned by the target LLM *Qwen3-4B-Base*. The results reveal a clear pattern: longer reasoning steps consistently receive higher step-level log probabilities, and a monotonic increasing relationship is observed between step length and log probability.

Short step, Low step probability:

***But again**, only some cases are covered.* **-6.69 -4.38** -2.46 -0.96 -1.29 -0.81 -0.11 -0.53	Step length: **8** Average prob: **-2.15**
***Wait,** but if \theta = 120, angle NBM = 30.* **-6.32** -1.47 -0.86 -0.69 -0.55 -1.02 -0.10 -0.10 ...	Step length: **13** Average prob: **-1.86**

Long step, High step probability:

***But again**, the problem is asking for a value of x, not necessarily the smallest or a specific one...* ***-5.48 -3.56** -1.43 -0.37 -0.16 -0.07 -0.25 -0.07 ...*	Step length: **67** Average prob: **-0.41**
***Wait**, hold on, in triangle ABC, the law of cosines says: cos \theta = (AB^2 + BC^2 − AC^2)/...* **-6.68** -0.82 -1.78 -1.70 -0.01 -0.05 -0.01 -0.70 ...	Step length: **58** Average prob: **-0.41**

Figure 3: Representative cases illustrating token-level log probabilities for varying step lengths.

For longer steps, the ratio of low-probability first tokens is lower. To further investigate the cause of the monotonic relationship between step length and step-level log probability, we examine several representative examples in Fig. 3, which illustrate short steps with low log probabilities and long steps with high log probabilities, respectively. Across all steps, *the first token consistently exhibits a lower log probability*. Previous studies have also confirmed this phenomenon (Wang et al., 2025; Cheng et al., 2025), attributing it to the fact that minority first tokens at each step often fork branches toward diverse reasoning pathways. Such branching behavior introduces higher entropy, which in turn yields lower log probabilities. Therefore, in longer steps, such low-probability first tokens always constitute a smaller proportion of the total tokens. Consequently, the larger number of non-first tokens dilutes the lower log probabilities of these first tokens, leading to a higher overall log probability and making such samples more likely to be selected by naturalness-based methods. In summary, our experiments lead to the following conclusion:

⋆ **Conclusion 2.** In naturalness-based methods, step length confounding occurs because **the low-probability first token constitutes a smaller ratio of longer responses**. This increases the average log probabilities, making samples with longer steps more likely to be selected.

Based on the above observations, we seek to design a debiasing approach targeting the first token to address the step length confounding problem.

[1] Notably, data selection actually also correlates with total response length (i.e., avg. L). However, as discussed in Appendix A, **the effect of total response length is substantially weaker than that of step length**.

3 The Proposed Method

In this section, we present our proposed variant methods ASLEC-DROP and ASLEC-CASL for LLM reasoning SFT data selection in detail.

Problem definition. Given i-th complex question $\mathbf{q}_i$ in the LLM reasoning SFT dataset $\mathcal{D}$, and K different responses $\mathbf{o}_i^k$ = <think> $\mathbf{c}_i^k$ </think> $\mathbf{a}_i^k$, $k \in \{1, \cdots, K\}$, where each $\mathbf{c}_i^k$ represents a reasoning trajectory and $\mathbf{a}_i^k$ the corresponding answer. These multiple responses may vary not only in correctness but also in reasoning quality, verbosity, and step length. Our method defines two metrics s_i^{drop} and s_i^{casl} to select one or more responses that best align with the trained reasoning LLM and are not confounded by the step length.

3.1 ASLEC-DROP: Dropping the First Token

As analyzed in Sec. 2.4, we attribute this step length confounding problem to the influence of the first token's probabilities (Bu et al., 2025). Consequently, the most straightforward approach is to drop the first token at each step when computing the geometric mean of the probabilities. Formally, we split a solution $\mathbf{o}_i$ into L reasoning steps $\mathcal{S}_i = \{\mathbf{s}_i^l\}_{l=1}^L$ and compute the metric as follows:

$$s_i^{\text{drop}} = \frac{1}{|\mathbf{o}_i| - |\mathcal{S}_i|} \sum_{\mathbf{s}_i^l \in \mathcal{S}_i} \sum_{t=2}^{|\mathbf{s}_i^l|} \log P_{\boldsymbol{\theta}} \left(s_{i,t}^l \mid \mathbf{s}_{i,<t}^l, \mathbf{s}_i^{<l}, \mathbf{q}_i \right), \tag{4}$$

where $\mathbf{s}_{i,<t}^l$ denotes the first t tokens of each step, and $\mathbf{s}_i^{<l}$ denotes all tokens across the first l steps.

3.2 ASLEC-CASL: Causally De-biasing

Although dropping the first token mitigates the bias it introduces, it simultaneously discards potentially informative signals carried by the first token itself. To address this trade-off, we draw inspiration from causal debiasing methods (Udomcharoenchaikit et al., 2022; Zhu et al., 2022), treating step length as a confounding factor and applying appropriate adjustments to account for its influence. To formalize this intuition, the log probability s_i^{logp} can be decomposed as the following linear regression equation:

$$s_i^{\text{logp}} = \beta_1 s_i^{\text{first}} + \beta_2 s_i^{\text{drop}} + \gamma \mathcal{Z}_i + \epsilon, \tag{5}$$

where s_i^{first} and s_i^{drop} represent the average log probabilities of the first token and the tokens excluding the first one in Eq. (4), respectively; $\mathcal{Z}_i$ represents the confounding factor, which in our method is defined as the proportion of the first token among all tokens; and ϵ denotes a residual noise term. The notation s_i^{first} and $\mathcal{Z}_i$ are formally given by:

$$s_i^{\text{first}} = \frac{1}{|\mathcal{S}_i|} \sum_{\mathbf{s}_i^l \in \mathcal{S}_i} \log P_{\boldsymbol{\theta}} \left(s_{i,1}^l \mid \mathbf{s}_i^{<l}, \mathbf{q}_i \right), \quad \mathcal{Z}_i = \frac{|\mathcal{S}_i|}{|\mathbf{o}_i|}, \tag{6}$$

where $s_{i,1}^l$ denotes the first token in the step $\mathbf{s}_i^l$. The basic idea of ASLEC-CASL is to adjust the raw log-probability score by removing the estimated influence of the confounder $\mathcal{Z}_i$. This yields a deconfounded metric s_i^{casl}, defined as:

$$s_i^{\text{casl}} \sim s_i^{\text{logp}} - \gamma \mathcal{Z}_i. \tag{7}$$

Accordingly, to calculate s_i^{casl} by estimating γ, given the dataset $\{s_i^{\text{logp}}, s_i^{\text{first}}, s_i^{\text{drop}}, \mathcal{Z}_i\}_{i=1}^N$, the parameters β_1, β_2, γ are estimated via ordinary least squares:

$$\min_{\beta_1, \beta_2, \gamma} \sum_{i=1}^{N} \left(s_i^{\text{logp}} - \beta_1 s_i^{\text{first}} - \beta_2 s_i^{\text{drop}} - \gamma \mathcal{Z}_i \right)^2. \tag{8}$$

This optimization admits a closed-form solution. Then the parameter vector is obtained as follows:

$$[\beta_1, \beta_2, \gamma]^\top = \left(\mathbf{X}^\top \mathbf{X} \right)^{-1} \mathbf{X}^\top \mathbf{Y}, \tag{9}$$

$$\mathbf{X}_{i,:} = \left[s_i^{\text{first}}, s_i^{\text{drop}}, \mathcal{Z}_i \right], \ \mathbf{Y}_{i,:} = \left[s_i^{\text{logp}} \right]. \tag{10}$$

Once γ is estimated, we compute the final deconfounded score $s_i^{\text{casl}} = s_i^{\text{logp}} - \gamma \mathcal{Z}_i$ for each instance and use it for downstream data selection.

4 Experimental Evaluation

In this section, we empirically evaluate the performance of our two proposed variant methods.

Evaluation settings. The experiments are conducted on two datasets, *LIMO-v2* and *AceReason-1.1-SFT*, using four different families of source LLMs: *QwQ-32B*, *Qwen3-32B*, *DeepSeek-R1-Distill-Qwen-32B*, and *gpt-oss-120b*, and four target LLMs of varying sizes: *Qwen3-4B-Base*, *Qwen3-8B-Base*, *Qwen3-4B-Instruct*, and *Qwen2.5-7B-Instruct*. Detailed descriptions of these LLMs and the implementation details of our SFT training are provided in Appendix B. We evaluate our trained LLMs on five benchmarks,

Table 1: Experimental results on *LIMO-v2* (Ye et al., 2025). We generate five responses per source LLM, and select five responses from these ones (select 4k responses from 16k data). The bold results represent the best scores.

Qwen	Method	*AIME24*		*AIME25*		*MATH500*	*OlympicB*	**Avg.**
		Accuracy	Pass@4	Accuracy	Pass@4			
4B-Base	+ GRACE (Zhang et al., 2025)	16.66	30.00	15.83	33.33	59.40	33.33	31.42
	+ Local LP (Just et al., 2025)	19.16	36.66	20.83	36.66	71.60	34.11	36.50
	+ ASLEC-DROP (ours)	30.00↑10.84	50.00↑13.34	28.33↑7.50	43.33↑6.67	77.80↑6.20	38.38↑4.27	44.64
	+ ASLEC-CASL (ours)	**31.66**↑12.50	**53.33**↑16.67	**30.83**↑10.00	**46.66**↑10.00	**80.00**↑8.40	**42.81**↑8.70	**47.54**
8B-Base	+ GRACE (Zhang et al., 2025)	30.83	53.33	21.66	36.66	72.00	39.70	42.36
	+ Local LP (Just et al., 2025)	34.16	53.33	20.83	36.66	76.60	42.81	44.06
	+ ASLEC-DROP (ours)	41.66↑10.50	**66.66**↑13.33	36.66↑15.83	43.33↑6.67	81.40↑4.80	47.85↑5.04	52.92
	+ ASLEC-CASL (ours)	**45.00**↑13.34	**66.66**↑13.33	**37.50**↑16.67	**53.33**↑16.67	**85.40**↑8.80	**49.03**↑6.22	**56.15**
4B-Instruct	+ GRACE (Zhang et al., 2025)	59.16	73.33	50.00	73.33	79.36	47.79	63.82
	+ Local LP (Just et al., 2025)	61.66	80.00	49.16	73.33	80.75	50.14	65.84
	+ ASLEC-DROP (ours)	69.16↑7.50	83.33↑3.33	56.66↑7.50	80.00↑6.67	89.88↑9.13	57.64↑7.50	72.77
	+ ASLEC-CASL (ours)	**71.66**↑10.00	**90.00**↑10.00	**58.33**↑9.17	**83.33**↑10.00	**93.20**↑12.45	**60.44**↑10.30	**76.16**
7B-Instruct	+ GRACE (Zhang et al., 2025)	17.50	26.66	11.66	23.33	61.50	32.35	28.83
	+ Local LP (Just et al., 2025)	17.50	30.00	10.83	26.66	64.68	33.97	30.60
	+ ASLEC-DROP (ours)	24.16↑6.66	40.00↑10.00	20.83↑10.00	36.66↑10.00	80.60↑15.92	41.17↑7.20	40.57
	+ ASLEC-CASL (ours)	**28.33**↑10.83	**46.66**↑16.66	**24.16**↑13.33	**46.66**↑20.00	**81.60**↑16.92	**45.18**↑11.21	**45.43**

including four mathematical reasoning datasets, *AIME24*, *AIME25*, *MATH500* (Lightman et al., 2023), and *OlympiadBench* (He et al., 2024), as well as one challenging scientific reasoning dataset *GPQA* (Rein et al., 2024). In addition, we compare two naturalness-based data selection methods: (i) GRAPE: selecting the responses with the highest s^{logp}; and (ii) Local LP: selecting the responses with the highest s^{loc}.

4.1 Main Results

Tables 1 and 3 present the experimental results of the four target LLMs on the *LIMO-v2* and *AceReason-1.1-SFT* datasets, respectively. Overall, both variants of our approach outperform the existing naturalness-based selection method, achieving average accuracy gains of 6.28% and 9.08% over the SOTA method Local LP (Just et al., 2025) on the two datasets. More specifically, prior naturalness-based methods, *e.g.,* GRACE and Local LP, are often hindered by the step length confounding problem, which leads them to overly prefer samples from a single data source. Consequently, their training performance consistently degrades and falls significantly below our method, which samples more evenly across diverse sources.

When comparing the two variant methods, ASLEC-CASL consistently outperforms ASLEC-DROP. This result suggests that the causal debiasing strategy successfully preserves the informative patterns embedded in the probability distribution of the first tokens. Meanwhile, the results indicate that our debiasing strategies are particularly

Table 2: Experimental performance on *GPQA*.

Method	Acc.	Pass@4	Acc.	Pass@4
	Qwen3-4B-Base		**Qwen3-4B-Insruct**	
GRACE	28.15	60.10	50.37	75.75
Local LP	29.16	61.61	52.14	77.27
ASLEC-DROP	34.97	65.65	58.83	83.33
ASLEC-CASL	**35.35**	**66.66**	**61.23**	**84.34**
Method	Acc.	Pass@4	Acc.	Pass@4
	Qwen3-8B-Base		**Qwen2.5-7B-Insruct**	
GRACE	47.97	75.25	25.37	56.06
Local LP	49.49	77.77	26.13	57.57
ASLEC-DROP	51.01	79.79	35.98	67.17
ASLEC-CASL	**52.14**	**82.32**	**38.51**	**74.74**

effective when data or model capacity is limited. For example, the performance gain on *LIMO-v2* exceeds that on *AceReason-1.1-SFT*, highlighting their strong suitability for low-resource SFT scenarios. In such settings, low-quality samples tend to have a more pronounced negative effect; our methods mitigate step length confounding and thereby enhance generalization performance.

We further compare our methods by training on *LIMO-v2* and evaluating on *GPQA*, a benchmark on the scientific domain. The experimental results again show that our approaches consistently and significantly outperform the SOTA Local LP selection baseline, and that the ASLEC-CASL variant achieves better performance than ASLEC-DROP.

4.2 Performance on Alleviating Confounding

To examine whether our proposed methods effectively address the step length confounding prob-

Table 3: Experimental results on *AceReason-1.1-SFT* (Chen et al., 2025b). We generate one response per source LLM, and select one response from these responses (select 10k responses from 40k data).

Qwen	Method	*AIME24*		*AIME25*		*MATH500*	*OlympicB*	**Avg.**
		Accuracy	Pass@4	Accuracy	Pass@4			
4B-Base	+ GRACE (Zhang et al., 2025)	30.00	60.00	29.16	36.66	77.20	42.50	43.82
	+ Local LP (Just et al., 2025)	32.50	63.33	30.00	40.00	77.80	43.08	45.22
	+ ASLEC-DROP (ours)	41.66↑9.16	**73.33**↑10.00	**30.83**↑0.83	43.33↑3.33	84.00↑6.20	47.20↑4.12	46.48
	+ ASLEC-CASL (ours)	**46.66**↑14.16	**73.33**↑10.00	**30.83**↑0.83	**46.66**↑6.66	**84.60**↑6.80	**48.23**↑5.15	**47.50**
8B-Base	+ GRACE (Zhang et al., 2025)	40.83	66.66	29.16	43.33	75.00	44.11	49.46
	+ Local LP (Just et al., 2025)	42.50	70.00	29.16	43.33	77.60	44.41	50.71
	+ ASLEC-DROP (ours)	50.00↑7.50	**76.66**↑6.66	36.66↑7.50	**53.33**↑10.00	86.20↑8.60	49.33↑4.92	54.60
	+ ASLEC-CASL (ours)	**53.33**↑10.83	**76.66**↑6.66	**39.16**↑10.00	**53.33**↑10.00	**86.60**↑9.00	**51.02**↑6.61	**56.21**
4B-Instruct	+ GRACE (Zhang et al., 2025)	54.16	73.33	43.33	63.33	84.80	48.82	59.65
	+ Local LP (Just et al., 2025)	57.50	76.66	42.50	66.66	85.60	49.26	61.79
	+ ASLEC-DROP (ours)	64.16↑6.66	**83.33**↑6.67	53.33↑10.83	76.66↑10.00	90.60↑5.00	57.18↑7.92	66.48
	+ ASLEC-CASL (ours)	**68.33**↑10.83	**83.33**↑6.67	**55.00**↑12.50	**80.00**↑13.34	**92.20**↑6.60	**57.64**↑8.38	**68.16**
7B-Instruct	+ GRACE (Zhang et al., 2025)	15.00	36.66	15.00	26.66	73.80	34.85	35.08
	+ Local LP (Just et al., 2025)	17.50	36.66	16.66	26.66	74.60	35.58	35.81
	+ ASLEC-DROP (ours)	25.00↑7.50	43.33↑6.67	22.50↑5.84	40.00↑13.34	82.00↑7.40	41.61↑6.03	42.35
	+ ASLEC-CASL (ours)	**30.00**↑12.50	**50.00**↑13.34	**24.16**↑7.50	**43.33**↑16.67	**82.40**↑7.80	**46.32**↑10.74	**46.07**

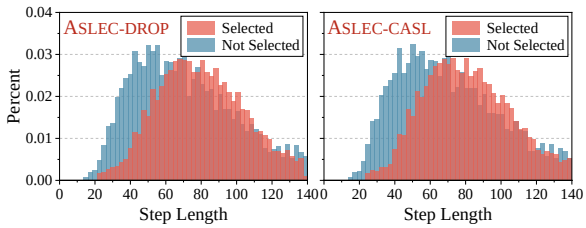

Figure 4: Step length distributions for data *selected* versus *unselected* by our two proposed variant methods.

lem, we present in Fig. 4 the distributions of step lengths for data samples *selected* and *unselected* by our two variants ASLEC-DROP and ASLEC-CASL. In contrast to the significant differences in step length distributions exhibited by prior approaches in Fig. 1, our approaches yield markedly smaller step length disparities. This demonstrates that our methods successfully mitigate step length confounding. Moreover, because both variants operate by intervening directly on the probability assigned to the first token, this result further implies that the step length confounding issue is intimately linked to the model's first-token probabilities.

4.3 Comparing Min and Max Probabilities

We also compare the performance of the ASLEC-CASL variant when selecting samples with the highest (max CASL) versus the lowest (min CASL) probability values of metric s^{casl}. All models are trained on the *LIMO-v2* dataset, and Table 4 reports their performance on four evaluation benchmarks after training the four target LLMs. The experimental results consistently show that selecting samples with the highest probabilities outperforms selecting the lowest ones, a finding aligned with previous naturalness-based selection methods. In general, high-probability samples correspond to data that better align with the target LLM's current capabilities, enabling more effective and stable learning of the SFT data distribution, thereby leading to superior performance. Conversely, lower-probability samples reflect that the model is less familiar with the data, which can introduce noisy training gradients and ultimately degrade model performance.

Table 4: Comparison of the experimental results for ASLEC-CASL that selects the highest and lowest s^{casl}.

Method	AIME24	AIME25	MATH	OlymB.
Qwen3-4B-Base				
max CASL	31.66	30.83	80.00	42.81
min CASL	29.16	28.33	77.40	39.70
Qwen3-8B-Base				
max CASL	45.00	37.50	85.40	49.03
min CASL	41.66	36.66	79.60	42.94
Qwen3-4B-Instruct				
max CASL	71.66	58.33	93.20	60.44
min CASL	65.83	55.83	86.80	55.14
Qwen2.5-7B-Instruct				
max CASL	28.33	24.16	81.60	45.18
min CASL	25.00	22.50	79.60	43.08

4.4 Linear Regression Results

Our method ASLEC-CASL fits a linear regression model as defined in Eq. (5), and uses this model to remove the influence of the first-token ratio on the average log probability. Accordingly, Table 5

Table 5: Linear regression parameters of Eq. (5) fitted on data generated by different source LLMs on *LIMO-v2*.

Source	β_1	β_2	γ	ϵ
QwQ-32B	0.077	0.919	-0.284	-0.001
Qwen3-32B	0.068	0.929	-0.529	0.007
DS-Qwen-32B	0.066	0.943	-0.226	0.001
gpt-oss-120b	0.068	0.938	**-1.284**	0.028
Overall	0.066	0.944	-0.680	-0.054

presents the fitted results of the linear regression for data originating from different source LLMs. First, γ is the most important coefficient, as it directly determines the extent to which the first-token ratio affects the average probability. Overall, the largest γ value among the models is -1.284, meaning that for every 0.05 difference in the first-token ratio between samples, the impact on the overall probability is equivalent to reducing each token's probability by $1 - e^{-1.284 \times 0.05} = 6.22\%$. For the regression fitted on all SFT data, γ is -0.680, corresponding to a $1 - e^{-0.680 \times 0.05} = 3.34\%$ reduction in per-token probability. Comparing different source LLMs, *gpt-oss-120b* exhibits the largest γ value, indicating that the data from it suffers from a more pronounced confounding problem.

In contrast, when comparing β_1 (the first-token probability) and β_2 (the non-first-token probability), we observe $\beta_1 \ll \beta_2$, which further suggests that the ratio of the first-token probability in the global average probability should be reduced. Lastly, ϵ consistently remains at a low level, implying that the regression model has only minor fitting bias, and thus the debiasing results are accurate.

4.5 Computation Budget

Compared to GRACE, our ASLEC-DROP variant adopts a more streamlined approach: it simply drops the first token when computing the average token probability, thereby avoiding any additional computational overhead. In contrast, our ASLEC-CASL variant introduces a modest amount of extra computation by fitting a lightweight linear regression model. However, since this regression model involves only a small number of parameters, the fitting procedure is highly efficient and typically completes in just a few seconds, imposing negligible cost on the overall pipeline.

5 Related Works

Recently, instead of direct prompt LLMs to generate CoT responses (Wei et al., 2022; Yuan et al., 2024; Bi et al., 2025), leveraging SFT to elicit long CoT reasoning in LLMs has emerged as a standard training paradigm, outperforming large-scale reinforcement learning even when applied to smaller models (Guo et al., 2025; Yang et al., 2025a; Tian et al., 2025; Kou et al., 2026). Generally, the existing methods typically scale up the SFT data generated by a strong LLM by constructing a large and diverse set of questions (Zhao et al., 2025; Guha et al., 2025; Yuan et al., 2025), and generating diverse solutions for each question through temperature sampling (Chen et al., 2023; Lei et al., 2025; Chen et al., 2025b; Yan et al., 2026). Building on these large-scale datasets, some studies have also sought to filter out noisy data by applying various heuristic rules, *e.g.,* question difficulty (Muennighoff et al., 2025; Guha et al., 2025; Li et al., 2025b), solution correctness (Chen et al., 2025a; Luo et al., 2025), diversity of solutions (Jung et al., 2025; Li et al., 2025a), and LLM-as-a-Judge (Wu et al., 2025; Lei et al., 2025). Remarkably, even reducing the dataset to only 1k questions can still elicit the long CoT reasoning capability (Muennighoff et al., 2025; Ye et al., 2025).

Beyond heuristic data-selection strategies, several works advocate naturalness-based approaches (Zhang et al., 2025; Just et al., 2025; Liu et al., 2026), wherein data are selected based on the model's confidence scores, allowing the selection of samples to which the model is better adapted. Although naturalness-based methods can indeed assess a model's adaptability to data via confidence scores (Kang et al., 2025; Fu et al., 2025), recent studies have shown that reasoning data often contain a small number of high-entropy / low-probability first tokens (Yang et al., 2025b; Wang et al., 2025). Our experiments further confirm that these low-probability first tokens can substantially exacerbate the step-length confounding problem in naturalness-based approaches.

6 Conclusion

In this work, we investigate the limitations of naturalness-based data selection for long CoT reasoning datasets. Our analysis reveals a systematic bias, termed step length confounding, whereby the selection pipeline significantly prefers samples with longer reasoning steps instead of those with higher reasoning quality. We trace this phenomenon to the disproportionate influence of low-probability first tokens in reasoning steps,

which is diluted in longer sequences, thus inflating their average log probabilities. To mitigate this problem, we propose ASLEC-DROP and ASLEC-CASL, two variants that drop or causally debias the first-token probability when computing selection scores. Extensive experiments on two reasoning SFT datasets, across four LLMs and five evaluation benchmarks, demonstrate that our approaches consistently outperform existing naturalness-based selection methods and effectively alleviate the step length confounding problem.

Limitations

This paper systematically investigates an important property in SFT data for LLM reasoning: the relationship between response and step length, and naturalness-based data selection. From a methodological perspective, one major limitation lies in our identification of the influence of a critical first token on step length confounding; whether there are deeper confounding factors remains an open question. In addition, recent studies have focused on on-policy data generation and selection (Yang et al., 2025a; Lu and Lab, 2025), where the student model produces its own samples for training. Whether data selection in such approaches still exhibits a strong correlation with response length is an issue worthy of further exploration.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (No.62276113).

References

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. *CoRR*, abs/2508.10925.

Jinhe Bi, Danqi Yan, Yifan Wang, Wenke Huang, Haokun Chen, Guancheng Wan, Mang Ye, Xun Xiao, Hinrich Schütze, Volker Tresp, and Yunpu Ma. 2025. Cot-kinetics: A theoretical modeling assessing LRM reasoning process. *CoRR*, abs/2505.13408.

Yuyan Bu, Liangyu Huo, Yi Jing, and Qing Yang. 2025. Beyond excess and deficiency: Adaptive length bias mitigation in reward models for RLHF. In *Findings of the Association for Computational Linguistics: NAACL*, pages 3091–3098.

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

Xiaoshu Chen, Sihang Zhou, Ke Liang, Xiaoyu Sun, and Xinwang Liu. 2025a. Skip-thinking: Chunk-wise chain-of-thought distillation enable smaller language models to reason better and faster. In *Conference on Empirical Methods in Natural Language Processing*, pages 12153–12168.

Yang Chen, Zhuolin Yang, Zihan Liu, Chankyu Lee, Peng Xu, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025b. Acereason-nemotron: Advancing math and code reasoning through reinforcement learning. *CoRR*, abs/2505.16400.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. 2025. Reasoning with exploration: An entropy perspective. *CoRR*, abs/2506.14758.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. 2025. The entropy mechanism of reinforcement learning for reasoning language models. *CoRR*, abs/2505.22617.

Yichao Fu, Xuewei Wang, Yuandong Tian, and Jiawei Zhao. 2025. Deep think with confidence. *CoRR*, abs/2508.15260.

Etash Kumar Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, and 1 others. 2025. Openthoughts: Data recipes for reasoning models. *CoRR*, abs/2506.04178.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, and 1 others. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Annual Meeting of the Association for Computational Linguistics*, pages 3828–3850.

Xingyue Huang, Rishabh, Gregor Franke, Ziyi Yang, Jiamu Bai, Weijie Bai, Jinhe Bi, Zifeng Ding, Yiqun Duan, Chengyu Fan, and 1 others. 2025. Loong: Synthesize long chain-of-thoughts at scale through verifiers. *CoRR*, abs/2509.03059.

Mingyu Jin, Qinkai Yu, Dong Shu, Haiyan Zhao, Wenyue Hua, Yanda Meng, Yongfeng Zhang, and Mengnan Du. 2024. The impact of reasoning step length on large language models. In *Findings of the Association for Computational Linguistics, ACL*, pages 1830–1842.

Jaehun Jung, Seungju Han, Ximing Lu, Skyler Hallinan, David Acuna, Shrimai Prabhumoye, Mostafa Patwary, Mohammad Shoeybi, Bryan Catanzaro, and Yejin Choi. 2025. Prismatic synthesis: Gradient-based data diversification boosts generalization in LLM reasoning. *CoRR*, abs/2505.20161.

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

Zhewei Kang, Xuandong Zhao, and Dawn Song. 2025. Scalable best-of-n selection for large language models via self-certainty. *CoRR*, abs/2502.18581.

Zhiqiang Kou, Junyang Chen, Xin-Qiang Cai, Xiaobo Xia, Ming-Kun Xie, Dong-Dong Wu, Biao Liu, Yuheng Jia, Xin Geng, Masashi Sugiyama, and Tat-Seng Chua. 2026. Positive-unlabeled reinforcement learning distillation for on-premise small models. *CoRR*, abs/2601.20687.

Zhenyu Lei, Zhen Tan, Song Wang, Yaochen Zhu, Zihan Chen, Yushun Dong, and Jundong Li. 2025. Learning from diverse reasoning paths with routing and collaboration. In *Conference on Empirical Methods in Natural Language Processing*, pages 2832–2845.

Hang Li, Kaiqi Yang, Yucheng Chu, Hui Liu, and Jiliang Tang. 2025a. Exploring solution divergence and its effect on large language model problem solving. *CoRR*, abs/2509.22480.

Yang Li, Youssef Emad, Karthik Padthe, Jack Lanchantin, Weizhe Yuan, Thao Nguyen, Jason Weston, Shang-Wen Li, Dong Wang, Ilia Kulikov, and Xian Li. 2025b. Naturalthoughts: Selecting and distilling reasoning traces for general reasoning tasks. *CoRR*, abs/2507.01921.

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

Kaiyuan Liu, Shaotian Yan, Rui Miao, Bing Wang, Chen Shen, Jun Zhang, and Jieping Ye. 2026. Where did this sentence come from? tracing provenance in LLM reasoning distillation. In *International Conference on Learning Representations*.

Kevin Lu and Thinking Machines Lab. 2025. On-policy distillation. *Thinking Machines Lab: Connectionism*.

Yijia Luo, Yulin Song, Xingyao Zhang, Jiaheng Liu, Weixun Wang, Gengru Chen, Wenbo Su, and Bo Zheng. 2025. Deconstructing long chain-of-thought: A structured reasoning optimization framework for long cot distillation. *CoRR*, abs/2503.16385.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A. Raffel. 2023. Scaling data-constrained language models. In *Annual Conference on Neural Information Processing Systems*.

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.

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 *Conference on Language Modeling*.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Na Zou, Hanjie Chen, and Xia Hu. 2025. Stop overthinking: A survey on efficient reasoning for large language models. *Transactions on Machine Learning Research*, 2025.

Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning.

Yijun Tian, Shaoyu Chen, Zhichao Xu, Yawei Wang, Jinhe Bi, Peng Han, and Wei Wang. 2025. Reinforcement mid-training. *CoRR*, abs/2509.24375.

Can Udomcharoenchaikit, Wuttikorn Ponwitayarat, Patomporn Payoungkhamdee, Kanruethai Masuk, Weerayut Buaphet, Ekapol Chuangsuwanich, and Sarana Nutanong. 2022. Mitigating spurious correlation in natural language understanding with counterfactual inference. In *Conference on Empirical Methods in Natural Language Processing*, pages 11308–11321.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. 2025. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. *CoRR*, abs/2506.01939.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

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. 2025. Beyond scaling law: A data-efficient distillation framework for reasoning. *CoRR*, abs/2508.09883.

Shaotian Yan, Kaiyuan Liu, Chen Shen, Bing Wang, Sinan Fan, Jun Zhang, Yue Wu, Zheng Wang, and Jieping Ye. 2026. Distribution-aligned sequence distillation for superior long-cot reasoning. *CoRR*, abs/2601.09088.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *CoRR*, abs/2505.09388.

Zhihe Yang, Xufang Luo, Zilong Wang, Dongqi Han, Zhiyuan He, Dongsheng Li, and Yunjian Xu. 2025b. Do not let low-probability tokens over-dominate in RL for llms. *CoRR*, abs/2505.12929.

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

Junjie Oscar Yin and Alexander M. Rush. 2025. Compute-constrained data selection. In *International Conference on Learning Representations*.

Weizhe Yuan, Jane Yu, Song Jiang, Karthik Padthe, Yang Li, Dong Wang, Ilia Kulikov, Kyunghyun Cho, Yuandong Tian, Jason E. Weston, and Xian Li. 2025. Naturalreasoning: Reasoning in the wild with 2.8m challenging questions. *CoRR*, abs/2502.13124.

Xiaosong Yuan, Chen Shen, Shaotian Yan, Kaiyuan Liu, Xiaofeng Zhang, Sinan Fan, Liang Xie, Wenxiao Wang, Renchu Guan, Ying Wang, and ieping Ye. 2026. Differential fine-tuning large language models towards better diverse reasoning abilities. In *International Conference on Learning Representations*.

Xiaosong Yuan, Chen Shen, Shaotian Yan, Xiaofeng Zhang, Liang Xie, Wenxiao Wang, Renchu Guan, Ying Wang, and Jieping Ye. 2024. Instance-adaptive zero-shot chain-of-thought prompting. In *Advances in Neural Information Processing Systems*.

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

Han Zhao, Haotian Wang, Yiping Peng, Sitong Zhao, Xiaoyu Tian, Shuaiting Chen, Yunjie Ji, and Xiangang Li. 2025. 1.4 million open-source distilled reasoning dataset to empower large language model training. *CoRR*, abs/2503.19633.

Congming Zheng, Jiachen Zhu, Zhuoying Ou, Yuxiang Chen, Kangning Zhang, Rong Shan, Zeyu Zheng, Mengyue Yang, Jianghao Lin, Yong Yu, and Weinan Zhang. 2025. A survey of process reward models: From outcome signals to process supervisions for large language models. *CoRR*, abs/2510.08049.

Yongchun Zhu, Qiang Sheng, Juan Cao, Shuokai Li, Danding Wang, and Fuzhen Zhuang. 2022. Generalizing to the future: Mitigating entity bias in fake news detection. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2120–2125.

A Bias Towards Response Length

In our preliminary experiments, we also observe that, under the same setup as in Fig. 1, the average total response length (*i.e.,* the total token number in the full response) of data *selected* by naturalness-based methods is approximately 9.8K, compared to about 15.4K for *unselected* data, revealing a significant discrepancy. Therefore, in the following sections, we aim to address the following question through experiments: *are samples with shorter overall response lengths more likely to be selected by naturalness-based data selection methods*?

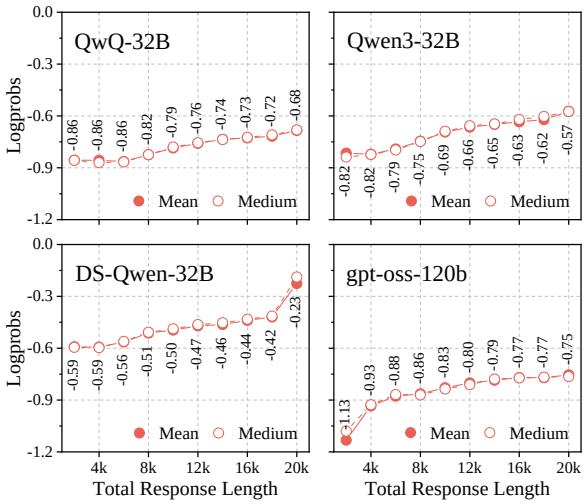

Figure 5: Relationship between response-level log probability and total response length.

A.1 Longer Response, Higher Log Probability

First, we maintain the same experimental setup as in Sec. 2.2, using *Qwen3-4B-Base* as the target model to compute the log probabilities for data generated by four different LLMs. Fig. 5 illustrates the trend of globally averaged log probabilities (*i.e.,* s_i^{logp}) *w.r.t* overall response length. The results show that, as the response length increases, the log probabilities actually rise, *contrary to the earlier conclusion, where shorter responses were more likely to be selected*. Comparing different models, we find that the average log probabilities of *DS-Qwen-32B*, which exhibits longer step lengths, are consistently and substantially higher than those of other models; even the highest log probability among other models is lower than the lowest value from *DS-Qwen-32B*. In summary, although log probabilities should in principle increase with response length, the selection bias caused by step length confounding has a much stronger influence than the effect of overall response length, leading to the seemingly opposite conclusion above.

Figure 6: Average log probability of tokens at different positions for four source LLMs.

A.2 Why Response Length Bias?

This section aims to further investigate why longer responses tend to have higher log probabilities. Fig. 6 presents the log probabilities of tokens at different positions within the responses. Specifically, we first determine the 95th percentile of the maximum response length, and then divide the range from 0 to this value into 20 bins. Tokens are assigned to these bins according to their position within the response, and the average log probability is computed for the tokens in each bin. As shown in Fig. 6, there is a clear trend: *tokens located toward the end of a response have higher log probabilities than those at the beginning*, following a monotonically increasing pattern. This is because, as the response length grows, the target LLM often becomes more confident in generating the continuation, indicating better adaptation to the data. Consequently, longer responses exhibit higher log probabilities for their tail-end tokens, which in turn results in a higher overall log probability.

A.3 Step Length Significantly Matters Response Length for Data Selection

In Sec. A.1, we present the experimental finding that the bias induced by total response length is negligible compared to the step length confounding issue. To further validate this conclusion, we employ the causal regression approach proposed in Sec. 3.2, rewriting Eq. (5) as follows:

$$s_i^{\text{logp}} = \beta_1 s_i^{\text{first}} + \beta_2 s_i^{\text{drop}} + \gamma \mathcal{Z}_i + \gamma_2 |\mathbf{o}_i| + \epsilon.$$

Using the same experimental settings, we refit the model, and the final parameter estimation results are reported in Table 6. Compared with the confounder $\gamma \mathcal{Z}_i$ introduced by step length, the confounder $\gamma_2 |\mathbf{o}_i|$ induced by total length is smaller by approximately two orders of magnitude.[2]

Table 6: Linear regression parameters including overall response length $\gamma_2 |\mathbf{o}_i|$.

Source	β_1	β_2	γ	γ_2	ϵ
QwQ-32B	.028	.972	-0.062	1E-8	.004
Qwen3-32B	.018	.979	-0.043	3E-9	.007
DS-Qwen-32B	.014	.983	-0.158	4E-8	-.008
GPT-OSS-120B	.018	.987	-1.166	5E-9	.025

Table 7: Comparison of experimental results with and without removing overall response length bias $\gamma_2 |\mathbf{o}_i|$.

Teacher	AIME24	AIME25	MATH	OlymB.
Qwen3-4B-Base				
ASLEC-CASL	31.66	30.83	80.00	42.81
+ $\gamma_2 \lvert\mathbf{o}_i\rvert$	30.83	30.83	78.60	42.20
Qwen3-8B-Base				
ASLEC-CASL	45.00	37.50	85.40	49.03
+ $\gamma_2 \lvert\mathbf{o}_i\rvert$	43.33	35.83	83.80	48.52

In Table 7, we further compare the performance metrics of models trained on data selected *with* and *without* the $\gamma_2 |\mathbf{o}_i|$ term in the criterion, *i.e.*, $s_i^{\text{casl}} = s_i^{\text{logp}} - \gamma \mathcal{Z}_i - \gamma_2 |\mathbf{o}_i|$. The results show that removing the $\gamma_2 |\mathbf{o}_i|$ term causes little change in model performance, once again confirming that the influence of total length is generally small and can even be negligible. In fact, prior studies have provided evidence that longer reasoning SFT data (Chen et al., 2025b; Guha et al., 2025) or in-context CoT prompts (Jin et al., 2024) can be more effective for improving model performance. This suggests that retaining the bias associated with total response length might even be beneficial. However, the step length confounding phenomenon leads to the opposite outcome, preferring shorter responses, which contradicts these findings and further underscores the importance of mitigating this bias.

B More Experimental Settings

In this section, we provide a detailed description of our experimental settings, including LLM model cards, data processing pipelines, and SFT details.

B.1 LLM Model Cards

In our experiments, we employ two categories of LLMs: those used to generate SFT data, which we refer to as *source LLMs*, and those trained on the generated SFT data, which we refer to as *target LLMs*. Their details are described as follows.

[2] $\mathcal{Z}_i$ is typically on the order of 10^{-1}, whereas $|\mathbf{o}_i|$ is typically on the order of 10^5.

Source LLMs. We use four different families of LLMs, each producing five distinct responses for every question.

- *QwQ-32B*[3] (Team, 2025) is a specialized reasoning model trained with reinforcement learning on top of *Qwen2.5-32B*.
- *Qwen3-32B*[4] (Yang et al., 2025a) has undergone large-scale long-CoT cold-start training and reasoning-focused reinforcement learning. In its technical report, this model is used to distill smaller-scale models, *e.g.,* the *Qwen3-4B* and *Qwen3-8B* variants, which align with our experimental setup.
- *DeepSeek-R1-Distill-Qwen-32B*[5] (Guo et al., 2025) is one of the first to employ reinforcement learning to enhance long CoT reasoning in LLMs, providing evidence that models obtained through distillation can still exhibit robust reasoning abilities.
- *gpt-oss-120b*[6] (Agarwal et al., 2025) improves inference speed by combining compact attention layers with linear attention layers, while activating only 5B parameters. The model is also trained using the conventional paradigm of SFT followed by reinforcement learning.

Target LLMs. Using the generated SFT data, we train four LLMs of varying sizes and types.

- *Qwen3-4B-Base*[7] and *Qwen3-8B-Base*[8] are two different-sized models that have undergone large-scale pre-training only.
- *Qwen3-4B-Instruct*[9] and *Qwen2.5-7B-Instruct*[10] build upon the two base versions described above, undergoing thorough instruction fine-tuning. For the 4B model, we adopt its 2507 variant, updating the non-thinking mode from *Qwen3-4B*'s mixed-reasoning framework. Furthermore, as the *Qwen3* series lacks an instruct model of 8B parameters, we use the 7B instruct model from the *Qwen2.5* series as a substitute.

[3] https://huggingface.co/Qwen/QwQ-32B
[4] https://huggingface.co/Qwen/Qwen3-32B
[5] https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-32B
[6] https://huggingface.co/openai/gpt-oss-120b
[7] https://huggingface.co/Qwen/Qwen3-4B-Base
[8] https://huggingface.co/Qwen/Qwen3-8B-Base
[9] https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507
[10] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

B.2 Data Sampling and Filtering

Our experiments are conducted using datasets from two different sources.

- *LIMO-v2*[11] (Ye et al., 2025) undergoes rigorous quality filtering, resulting in a final selection of 800 high-quality mathematics problems. For this dataset, we generate five diverse correct responses for each problem using every source LLM.
- *AceReason-1.1-SFT*[12] (Chen et al., 2025b) aggregates large-scale, high-quality SFT data from multiple sources. From this dataset, we randomly sample 10k mathematics problems, and for each problem, we obtain one correct response generated by each of the four aforementioned source LLMs.

For these two datasets, we adopt the following data sampling and quality filtering pipeline.

Data sampling. During data sampling, we employ top-p sampling with p set to 0.95. For gpt-oss-120b, following the official recommendations, we set the sampling temperature to 1.0 and the reasoning effort to high. For the other source LLMs, the sampling temperature is set to 0.6. All data sampling is conducted via SGLang for offline LLM deployment and calling, with the maximum generation length consistently fixed at 64K tokens.

Data filtering. We perform dynamic filtering during the sampling process. Specifically, for each problem, we sample one response at a time and verify it using the Math-Verify toolkit.[13] Sampling continues until the required number of correct responses is obtained (5 responses for *LIMO-v2* and 1 response for *AceReason-1.1-SFT*), or until the number of attempts exceeds 15. In practice, some problems are difficult to collect five correct responses for, so we repeat the above procedure five times to ensure that every problem has sufficient correct responses. For problems that are excessively difficult and fail to meet the required number of correct responses, we adopt the following remedies. In *LIMO-v2*, we manually sample several responses that are close to the correct answer and supplement them until five correct responses are obtained. In *AceReason-1.1-SFT*, we sample additional problems from the dataset and continue generation until

[11] https://huggingface.co/datasets/GAIR/LIMO-v2
[12] https://huggingface.co/datasets/nvidia/AceReason-1.1-SFT
[13] https://github.com/huggingface/Math-Verify

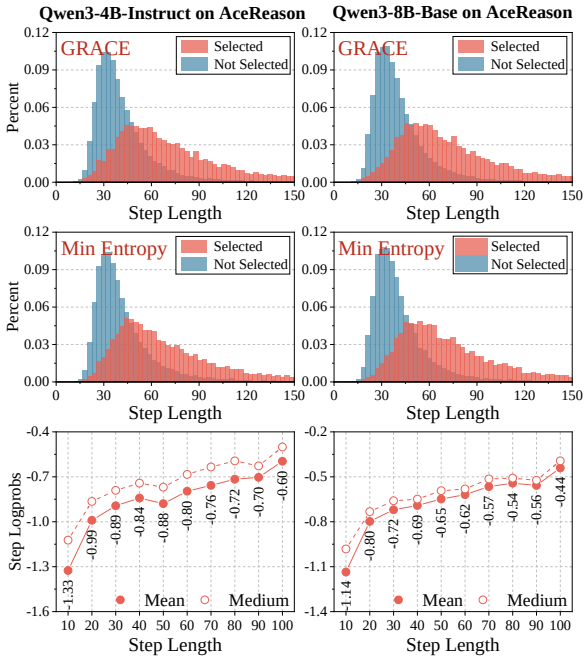

Figure 7: Step length distributions for *selected* and *unselected* data and relationship between step-level log probability and step length on *AceReason-1.1-SFT*.

we reach a total of 10k problems, each paired with its corresponding correct response. As a result, each problem in *LIMO-v2* may contain up to 75 incorrect responses and at least 25 correct ones. All generated response data are publicly available via the link provided in this paper.

Sampling log probabilities of target LLMs. We employ SGLang to sample log probabilities from the target LLM for the data. In GRACE (Zhang et al., 2025), the output log probabilities are averaged directly. In Local LP (Just et al., 2025), following the best practices outlined in the original paper, each problem and step is paired with its preceding $k = 4$ steps, the average log probability is computed for each step, and the step-level averages are then averaged. For Min Entropy, retrieving vocabulary-level probability distributions for every token is computationally and storage-intensive. Given their typical long-tailed nature, we instead retain only the top 20 most probable tokens per token, compute the entropy for each, and average these values across all tokens in a response.

B.3 Fine-tuning Details

Using the reasoning SFT data, we fine-tune the target LLM with full-parameter training via LlamaFactory. We set the training batch size to 32 and enable LlamaFactory's built-in *packing* option to concatenate shorter samples, with the maximum sequence length fixed at 32K. Optimization is carried out using the Adam optimizer with a learning rate of 5×10^{-5} for a total of 6 epochs. The learning rate scheduler is configured as *cosine_with_min_lr*, with a minimum learning rate of 1×10^{-5}.

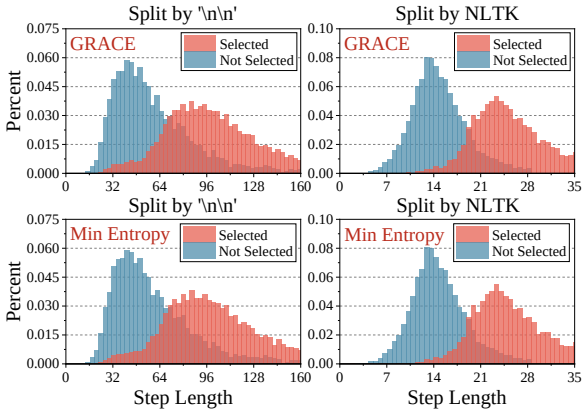

Figure 8: Data selection bias and step length distributions under different splitting methods.

C More Analysis Results

In this section, we provide additional analytical results regarding step length confounding.

C.1 Analysis on More Datasets and Models

In Sec. 2, we analyze the *LIMO-v2* dataset using *Qwen3-4B-Base* as the target LLM, examining results across four different source LLMs. In this section, we extend our analysis to the *AceReason-1.1-SFT* dataset and combine data from all source LLMs, evaluating their performance on two target models: *Qwen3-4B-Instruct* and *Qwen3-8B-Base*. The analysis results are shown in Fig. 7. The experimental results show that, for both target LLMs, the *AceReason-1.1-SFT* dataset exhibits a pronounced difference in step-length distribution between the *selected* and *unselected* samples. Moreover, the monotonic increasing relationship between each step's log probability and its length remains significant, with the probabilities from *Qwen3-8B-Base* consistently exceeding those of *Qwen3-4B-Instruct*.

C.2 Different Step Splitting Methods

The step-length distribution differences shown in Fig. 1 are obtained by splitting the responses into steps using periods and spaces. In the community, aside from this splitting approach, some studies adopt \n\n or external tools such as NLTK for step splitting. Therefore, we also investigate the

Table 8: Linear regression parameters of all SFT data on *AceReason-1.1-SFT* for the two target models.

Target LLM	β_1	β_2	γ	ϵ
Qwen3-4B-Instruct	0.075	0.952	-0.778	0.022
Qwen3-8B-Base	0.027	0.993	-0.660	0.025

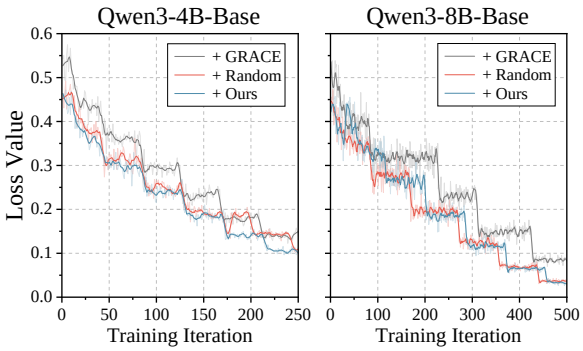

Figure 9: Convergence analysis.

step length-distribution differences under these two additional step splitting methods. The analysis results are presented in Fig. 8.

Our experimental results consistently show that, regardless of the step splitting method used, the *selected* versus *unselected* data still display a clear difference in step-length distribution. This indicates that the naturalness-based data selection approach continues to suffer from a step length confounding issue. Moreover, splitting sentences using periods produces the most distinct distribution differences compared to other splitting methods. This leads to another key observation: low-probability tokens are most prominent at sentence beginnings when period-based splitting is applied, as opposed to the first tokens in steps segmented by other methods.

C.3 Casual Regression Parameters

We apply the causal regression method introduced in Sec. 3.2 to the *AceReason-1.1-SFT* dataset, re-fitting the two target LLMs analyzed previously, and present the regression parameters in Table 8. The results show that the γ values remain high, indicating that the step length confounding issue persists. Furthermore, the result $\beta_1 \ll \beta_2$ is fully consistent with the conclusions presented in Sec. 4.4.

D More Experimental Results

D.1 Convergence Analysis

In Fig. 9, we present a convergence analysis showing that the GRACE method, *i.e.,* the existing naturalness-based approach, consistently converges to a higher loss compared with our method. This also demonstrates that our debiasing approach is able to select data with greater naturalness.

Table 9: Results of more data selection methods.

Qwen3-4B-Base	AIME25	MATH500
+ Uniform	24.16	73.20
+ High Difficulty	26.66	77.80
+ Low Difficulty	20.83	71.20
+ ASLEC-CASL	**30.83**	**80.00**

Table 10: Results on Llama3-3B.

Llama3-3B	AIME25	MATH500
+ GRACE	26.66	76.20
+ ASLEC-CASL	**33.33**	**84.40**

D.2 Comparing More Baselines

We implement a simple heuristic-based selection method on the same source data as a representative baseline. The results of this comparison are presented in Table 9. On the LIMO-v2 dataset, we compare several baseline selection strategies:

- **Uniform**: randomly samples 4k reasoning trajectories from the four source LLMs with a uniform distribution to ensure data diversity;
- **High/Low Difficulty**: selects 4k of the longest or shortest reasoning trajectories, respectively, using response length as a proxy for problem difficulty.

The results consistently demonstrate the superior effectiveness of our selected data. Moreover, while more difficult (*i.e.,* longer) examples do yield better performance for SFT of reasoning models, confirming that challenging instances are generally more informative, they still contain a higher proportion of redundant or noisy reasoning steps compared to our method. This underscores the advantage of our approach in not only capturing useful difficulty but also filtering out superfluous or low-quality reasoning content.

D.3 Comparing More Source LLMs

To further assess generalizability, we also train a non-Qwen model: Llama3-3B, on the data selected by our method ASLEC-CASL; the results are presented in Table 10. The results consistently demonstrate the effectiveness of our proposed method.