NATURALTHOUGHTS: Selecting and Distilling Reasoning Traces for General Reasoning Tasks

Yang Li[*] Meta
Youssef Emad[*] Meta
Karthik Padthe[*] Meta
Jack Lanchantin[*] Meta
Weizhe Yuan Meta
Thao Nguyen Meta
Jason Weston Meta
Shang-Wen Li Meta
Dong Wang Meta
Ilia Kulikov[†] Meta
Xian Li[†] Meta

Abstract

Recent work has shown that distilling reasoning traces from a larger teacher model via supervised finetuning outperforms reinforcement learning with the smaller student model alone (Guo et al., 2025). However, there has not been a systematic study of what kind of reasoning demonstrations from the teacher are most effective in improving the student model's reasoning capabilities. In this work we curate high-quality NATURALTHOUGHTS by selecting reasoning traces from a strong teacher model based on a large pool of questions from NATURALREASONING (Yuan et al., 2025). We first conduct a systematic analysis of factors that affect distilling reasoning capabilities, in terms of sample efficiency and scalability for general reasoning tasks. We observe that simply scaling up data size with random sampling is a strong baseline with steady performance gains. Further, we find that selecting difficult examples that require more diverse reasoning strategies is more sample-efficient to transfer the teacher model's reasoning skills. Evaluated on both Llama and Qwen models, training with NATURALTHOUGHTS outperforms existing reasoning datasets such as OpenThoughts, LIMO, etc. on general STEM reasoning benchmarks including GPQA-Diamond, MMLU-Pro and SuperGPQA.

1 Introduction

Scaling reinforcement learning (RL) with generated chain-of-thoughts (CoTs) has led to remarkable improvements in the reasoning capabilities of large language models (LLMs) (Guo et al., 2025). An effective approach to elicit such capabilities, especially for smaller models, is to distill from a teacher model, i.e. supervised-finetuning (SFT) on reasoning traces output by a stronger reasoning model. In practice, SFT with thousands of training examples is often applied as a critical step before RL even for larger models (Guo et al., 2025; Bercovich et al., 2025). The importance of distillation is also evident by the fact that RL alone does not increase the innate priors for reasoning a student model (Yue et al., 2025), while SFT on reasoning traces from a teacher model can add new reasoning primitives to be explored in the RL stage.

Various efforts from the community have been devoted to distilling CoT trajectories from reasoning models, with notable examples being Open-R1 (Face, 2025) and OpenThoughts (Guha et al., 2025). Through these efforts, it has been demonstrated that state-of-the-art reasoning models often exhibit sub-optimal reasoning trajectories such as "overthinking" or "underthinking" (Kumar et al., 2025; Chen et al., 2025; Wang et al., 2025). Recent work such as LIMO (Ye et al., 2025) and S1K (Muennighoff et al., 2025) show the importance of selecting and curating distillation examples to improve downstream capabilities. They find that as few as 1,000 high-quality reasoning traces are sufficient to drastically increase the student model's performance on mathematical reasoning tasks (Ye

[*]Joint first author, [†]Joint last author

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Efficient Reasoning.

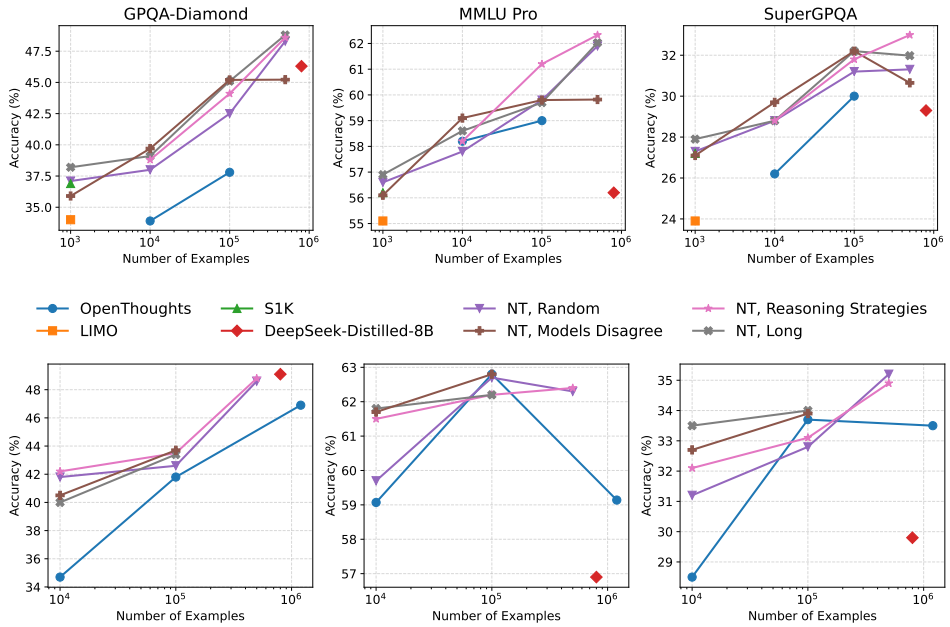

Figure 1: Comparison of NATURALTHOUGHTS (NT) with existing distillation datasets, when training Llama-3.1-8B-Instruct (Top) and Qwen-2.5-7B-Instruct (Bottom) respectively. In contrast to the "Less is More" hypothesis (Ye et al., 2025; Muennighoff et al., 2025), we observe that scaling up high-quality questions and reasoning samples consistently improves performance, even with random selection. Selection based on reasoning strategies diversity and difficulty (e.g. long CoT, disagreement between teacher models) further improves over random selection (Section 2.2, Section 4.1).

et al., 2025; Muennighoff et al., 2025). Both LIMO and S1K conduct manual selection of distillation examples, limited to math and coding questions, subsequently leading primarily to performance gains in those domains. While such curated small-scale datasets are helpful for solving easy to medium-difficulty problems in in-distribution domains, they may not generalize well to reasoning problems in other domains (Sun et al., 2025).

In this work, we conduct a systematic analysis of data-centric factors that affect distilling reasoning capabilities from a "reasoning" teacher model to an (initially) "non-reasoning" student model. Specifically, we use questions from NaturalReasoning as a testbed, given its diversity and effectiveness in eliciting reasoning (Yuan et al., 2025), and generate reasoning traces from a performant reasoning model such as DeepSeek-R1 (Guo et al., 2025). We use the resulting dataset, NATURALTHOUGHTS, to examine the effects of training on different filtered data subsets and understand how models learn to reason effectively. We study the curation of distillation training examples along several axes including scale, diversity, and difficulty. We find that increasing each axis in the training data leads to performance gains.

Given that an important consideration when deploying the student model in real-world applications is inference-time efficiency, we also take into account the susceptibility of current models to overthinking (i.e., generating too many reasoning tokens). Consequently, in addition to reasoning performance, we also focus on improving the student model's reasoning *efficiency*. We propose a simple training method with mixed System-1 and System-2 distillation (Yu et al., 2024a), where training examples do not always contain the full reasoning traces from the teacher (System-2), but instead only contain the teacher model's final answers after thinking (System-1).

Our main contributions are as follows:

- We demonstrate the importance of **scaling** high-quality, diverse reasoning data, which is contrary to the "Less is More" hypothesis (Ye et al., 2025; Muennighoff et al., 2025) (Section 4.1). We observe consistent performance improvements by scaling up the data quantity, even with *random* selection from NATURALTHOUGHTS.

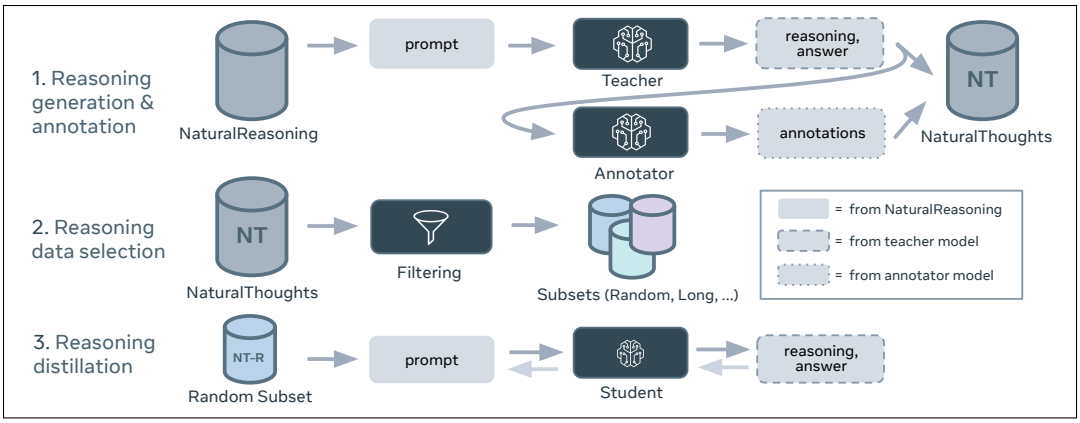

Figure 2: NATURALTHOUGHTS Overview. We outline the three main components of our strategies for selecting and distilling reasoning data. 1) Starting from the NaturalReasoning dataset (Yuan et al., 2025), we sample prompts for the teacher model to generate reasoning and answer traces from, creating the NATURALTHOUGHTS dataset. 2) From the NATURALTHOUGHTS seed set, we use several filtering methods for data selection (Section 4.1). 3) Given a filtered subset obtained from data selection, we finetune a student model on the prompt-reasoning-answer triples, where the reasoning traces and answers are generated by the teacher model.

- We systematically compare different data **selection** methods based on a large pool of reasoning traces from a teacher model (Section 4.1). We examine the efficacy of selecting SFT examples based on various metrics including *diversity* and *difficulty*. We find that difficult examples which require longer reasoning chains and diverse reasoning strategies are most effective at distilling the teacher's reasoning capabilities.
- To make the student model's reasoning process more **efficient**, we show that a simple method of training with mixed System-1 (using the condensed final responses) and System-2 (using both the intermediate CoTs and the final responses) reasoning based on question difficulty enables the student model to adapt its reasoning strategy based on the input. This allows for a dynamic trade-off between efficiency and accuracy at test time, resulting in more effective problem-solving and significantly shifting the efficiency-accuracy frontier. (Section 4.3).

2 Method

The objective of our work is to distill reasoning capabilities from a "reasoning" teacher model to an (initially) "non-reasoning" student model. We define each reasoning training example to be a (question, reasoning response) pair, where the response is generated by a teacher model, consisting of two parts: the intermediate reasoning trace (e.g. the tokens between `<think>` and `</think>`, which represent *System-2* reasoning) and the final answer (*System-1*). We sample questions from NaturalReasoning (Yuan et al., 2025), a comprehensive dataset comprising 2.8 million questions that span multiple domains and that have been shown to be effective in eliciting reasoning. We use DeepSeek-R1 (Guo et al., 2025) as the teacher model to generate reasoning responses. We call the resulting dataset of distilled reasoning examples NATURALTHOUGHTS. In the following subsections, we outline how we annotate and filter reasoning examples from the initial dataset in order to optimize the student model's reasoning capabilities.

2.1 Reasoning Annotation

Given a training set of (question, reasoning response) pairs, we first seek to annotate the data, using three different annotations. First, we annotate the domain and topics of each question using the taxonomy from Du et al. (2025), which includes 13 top-level domains including Engineering, Philosophy, Medicine, Economics, Science, Law, History, Education, Management, Literature and Arts, Agronomy, Sociology and Military Science. Then, for each reasoning trace, we prompt Llama-3.1-70B-Instruct (Grattafiori et al., 2024) to identify the "meta-reasoning" strategies throughout the

thinking process, such as self-verification, backtracking, exploration, etc. Finally, we prompt Llama-3.1-70B-Instruct to score the "verbosity" of the reasoning, from 0 to 10, where 0 means the reasoning is very efficient with no rambling, and 10 means excessive rambling without making progress towards a solution. Full prompts used for annotations are provided in Figure 9 and Figure 10.

2.2 Reasoning Data Selection

Given the training examples with their respective annotations, the second step is to select samples used for distillation. We study different data selection strategies along the following two axes: diversity and difficulty.

Diversity We hypothesize that a diversified set of questions and reasoning traces would be effective for distillation (Muennighoff et al., 2025). We attempt to obtain diverse subsets of data using three properties.

- **Question topics** With annotated question topics from Section 2.1, we sample data uniformly across all the 12 topic domains, taking about 850 samples from each domain.
- **Question semantic embeddings** We also create a diverse subset of questions using the semantic feature space. Specifically, we embed questions using Llama 3.1-8B-Instruct. We then perform density-based clustering on the embeddings, and uniformly sample from each cluster. More details can be found in Appendix D.
- **Reasoning strategies** With the annotated reasoning strategies from Section 2.1, each example has a set of strategies $\mathcal{S} = \{s_i\}$. As shown in Figure 4, there are a few frequently used strategies such as self-verification, etc. followed by a long tail of niche strategies. To select examples demonstrating diverse problem-solving strategies yet without "overthinking", we downsample examples where the number of reasoning strategies $|S| \leq R_{min}$ or $|S| > R_{max}$. In the experiments in Section 4.1, we use $R_{min} = 4$ and $R_{max} = 8$, based on the distribution of unique strategies annotated per example (Figure 8). We also downsample examples with low reasoning density, measured as having fewer unique reasoning strategies than the number of reasoning steps. Further ablations on the sampling approach are provided in Appendix E.

Difficulty Another hypothesis is that the quality of the reasoning traces is correlated with the difficulty of the questions, which usually requires advanced reasoning. We therefore attempt to create data subsets with varying levels of difficulty using the following strategies.

- **Length** The initial training dataset is dominated by short traces. Therefore, we downsample examples with short reasoning responses to study the benefit of using longer reasoning chains. Specifically, each example is sampled with probability $p = (l/C)^{\tau}$, where l is the length of reasoning response measured by the number of tokens, C is a constant normalizer, and τ is the sampling temperature modulating how heavily shorter sequences are downsampled. In the experiments in Section 4.1, we use $C = 5000$, $\tau = 2.5$.
- **Verbosity** Given the annotated verbosity scores from Section 2.1, we derive three subsets by sampling without replacement based on the scores: Low (beginning with the lowest verbosity, 0, and progressively including samples with higher verbosity), High (starting from the highest verbosity, 10, and progressively including samples with lower verbosity), and Med (including all samples with a verbosity of 5).
- **Models Agreement** For each example, we compared responses from a model with long CoT reasoning traces (Deepseek-R1) and a model without long CoT traces (Llama-3.3-70B). We use their disagreement, judged by Llama-3.1-8B-Instruct, as a proxy of question difficulty. We create two subsets of training examples based on solution agreement or disagreement.

2.3 Mixed Reasoning Distillation

As the teacher reasoning model may have sub-optimal reasoning patterns such as "overthinking" or "underthinking", we compare different settings of distilling the teacher model's reasoning:

System-2 Distillation By default, we conduct supervised finetuning on the entire response generated by the teacher model, which includes the full reasoning trace and the final answer.

System-1 Distillation Instead of learning from the long CoT reasoning trace, we investigate the effectiveness of only learning from the teacher's final answer.

Mixed System-1 and System-2 Training data is a mixture of examples from both types as described above. We compared two mixing approaches: *random mixing* and *difficulty-based mixing*. In random mixing, we select training examples with full System-2 reasoning with probability $p = \{0.2, 0.4, 0.6\}$. In difficulty-based mixing, we use full System-2 reasoning traces for examples annotated with disagreement (as a proxy of difficult questions) and only use the condensed System-1 response for the remaining examples.

Adaptive Reasoning at Inference-time To enable explicit control of which types of reasoning to use at inference time, we append an explicit instruction to the end of the question to indicate which types of reasoning and how much inference budget the response should use. Specifically, we augment the training data with "`Think carefully before answering. Use about {K} words.`" for System-2, and "`Answer directly without thinking. Use about {K} words.`" for System-1. `K` is derived from the training data. At inference time, we evaluate the accuracy-efficiency trade-offs under three settings:

- **No-Think** We instruct the model to "`Answer directly without thinking`", i.e. perform System-1 mode of reasoning by generating a short condensed answers.
- **Think** We instruct the model to "`Think carefully before answering. Use about {K} words`" followed by the special think start token `<think>` to force the generation into System-2 mode. K is set to 3500 which is the average length of System-2 responses in the training set.
- **Adaptive-Think** To test whether the mixed System-1 and System-2 distillation can enable the student model to efficiently and automatically adapt to the question difficulty at inference time, we also evaluate a *hybrid* mode, where we instruct the model to "Think carefully before answering." but *without* explicitly appending the special token `<think>`.

3 Experimental Setup

We perform supervised finetuning (SFT) with NATURALTHOUGHTS data on both Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct student models. We compare different data selection methods at the scale of 1,000 and 10,000 training examples. We also experiment with training on 100,000 and 500,000 samples to test the scaling properties of NATURALTHOUGHTS. During SFT training, we mask the loss on the question, i.e. only train on the teacher model's reasoning trace and final answer. Unless otherwise noted, we use a maximum response length of 16,384 tokens and ensure that each training example contains a complete response within the maximum number of tokens. We utilize the fairseq2 library (Balioglu et al., 2023) for training. We apply dynamic batching where each batch roughly contains 400k tokens. We train 10 epochs for 1k samples, 6 epochs for 10k, and 8 epochs for 100k and 500k training examples respectively. We use the AdamW optimizer with 0.1 weight decay and a constant learning rate $2e-5$. Each model is trained on 32 NVIDIA H200 GPUs.

Baselines We compare the curation approaches described in Section 2.2 with the following baselines:

- **Random selection** We randomly select 1k, 10k, 100k and 500k examples from NATURALTHOUGHTS (NT-RANDOM).
- **State-of-the-art reasoning datasets** We compare to LIMO (Ye et al., 2025) and S1K (Muennighoff et al., 2025) which use the same teacher model (DeepSeek-R1) and perform careful manual curation for both the questions and the reasoning responses. For a fair comparison, we train our own Llama-3.1-8B-Instruct student model on their datasets. To understand the effect of question source and question quality in NATURALTHOUGHTS, we also build another two baselines (NT-NN-LIMO, NT-NN-S1K), which use questions from LIMO and S1K as seeds to retrieve similar questions from NATURALTHOUGHTS. Specifically, we choose the nearest neighbors of the LIMO and S1K questions in the question embedding space, and train a Llama-3.1-8B-Instruct student model on the retrieved data. We also compare to OPENTHOUGHTS3 (Guha et al., 2025), a high-quality reasoning dataset that has achieved state-of-the-art performance on several reasoning benchmarks. We compare NATURALTHOUGHTS to OPENTHOUGHTS3 using both Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct student models[2].

[2] Specifically, we compare to different released assets from OPENTHOUGHTS3 for different settings. For experiments with Llama models, as described in Table 1, we compare to https://huggingface.co/mlfoundations-dev/openthoughts3_10k_llama3 and https://huggingface.co/mlfoundations-dev/openthoughts3_100k_llama3. For experiments with Qwen models, as described in Table 4, we compare to https://huggingface.co/mlfoundations-dev/openthoughts3_10k, https://huggingface.co/mlfoundations-dev/openthoughts3_100k and https://huggingface.co/open-thoughts/OpenThinker3-7B.

- **DeepSeek-R1 distilled models** As a reference point, we also compare to the DeepSeek-R1-Distill-Llama-8B and DeepSeek-R1-Distill-Qwen-7B (Guo et al., 2025) models, which were trained via SFT with the same teacher and student models as our experiments, but with 800k non-public data.

Evaluation We evaluate on reasoning benchmarks in both math and general STEM domains: MATH-500 (Hendrycks et al., 2021) (500 examples), GPQA-Diamond (Rein et al., 2024) (198 examples), MMLU-Pro (Wang et al., 2024) (12,032 examples), and SuperGPQA (Du et al., 2025) (26,529 examples). We evaluate all models using the same evaluation setting as used in Guo et al. (2025). Specifically, we report $pass@1$ averaged across multiple seeds to reduce the variance. We use $N = 24$ seeds for GPQA-Diamond, and $N = 16$ for MATH-500, and N=1 for MMLU-Pro and SuperGPQA. For decoding hyperparameters, we use temperature 0.6, and top-p 0.95 for Llama models, and temperature 0.7, top-p 1.0 for Qwen models. We use maximum generation length of 16,384 tokens to match the sequence lengths used in training.

4 Results

	Train Size	GPQA-D	MATH-500	MMLU-Pro	SuperGPQA
Baselines					
Llama-3.1-8B-Instruct	0	$29.0_{\pm 0.5}$	$49.1_{\pm 0.4}$	47.7	21.9
LIMO (Ye et al., 2025)	817	$34.0_{\pm 0.5}$	$56.5_{\pm 0.3}$	55.1	23.9
S1K (Muennighoff et al., 2025)	1k	$36.9_{\pm 0.7}$	$59.4_{\pm 0.3}$	56.2	27.2
NT-NN-LIMO	10k	$35.3_{\pm 0.6}$	$64.7_{\pm 0.4}$	57.6	27.7
NT-NN-S1K	10k	$38.5_{\pm 0.6}$	$63.7_{\pm 0.3}$	57.4	29.0
OpenThoughts3 (Guha et al., 2025)	10k	$33.9_{\pm 0.5}$	$72.3_{\pm 0.4}$	58.2	26.2
OpenThoughts3 (Guha et al., 2025)	100k	$37.8_{\pm 0.5}$	$82.2_{\pm 0.2}$	59.0	30.0
DeepSeek-R1-Distill-Llama-8B	800k	46.3	89.1	56.2	29.3
NATURALTHOUGHTS, *Scale*					
Random	1k	$37.1_{\pm 0.4}$	$57.8_{\pm 0.3}$	56.6	27.3
Random	10k	$37.6_{\pm 0.6}$	$61.3_{\pm 0.3}$	57.8	28.8
Random	100k	$42.5_{\pm 0.6}$	$67.5_{\pm 0.3}$	59.8	31.2
Reasoning Strategies	100k	$44.1_{\pm 0.6}$	$67.7_{\pm 0.3}$	61.2	31.8
Models Disagree	100k	$45.2_{\pm 0.6}$	$70.2_{\pm 0.2}$	59.8	32.2
Long	100k	$43.1_{\pm 0.7}$	$69.0_{\pm 0.3}$	61.2	32.2
Random	500k	$48.3_{\pm 0.6}$	$72.3_{\pm 0.3}$	61.9	31.3
Reasoning Strategies	500k	$48.6_{\pm 0.7}$	$75.4_{\pm 0.2}$	62.3	33.0
Models Disagree	500k	$45.2_{\pm 0.6}$	$70.8_{\pm 0.3}$	59.8	30.7
Long	500k	$48.8_{\pm 0.6}$	$74.9_{\pm 0.3}$	62.0	31.9
NATURALTHOUGHTS, *Diversity*					
Topics	10k	$32.7_{\pm 0.5}$	$55.7_{\pm 0.3}$	55.9	25.4
Semantic Embeddings	10k	$39.4_{\pm 0.5}$	$60.3_{\pm 0.3}$	55.9	27.2
Reasoning Strategies	10k	$38.8_{\pm 0.5}$	$63.5_{\pm 0.4}$	58.2	28.8
NATURALTHOUGHTS, *Difficulty*					
Long	10k	$39.1_{\pm 0.7}$	$63.5_{\pm 0.3}$	58.6	28.8
Models Disagree	10k	$39.7_{\pm 0.4}$	$61.9_{\pm 0.3}$	59.1	29.7
Models Agree	10k	$37.5_{\pm 0.6}$	$60.1_{\pm 0.3}$	56.7	28.9
Verbosity=Low	10k	$37.0_{\pm 0.7}$	$59.1_{\pm 0.3}$	54.8	26.7
Verbosity=Med	10k	$37.2_{\pm 0.6}$	$62.1_{\pm 0.4}$	57.5	28.9
Verbosity=High	10k	$38.4_{\pm 0.6}$	$59.8_{\pm 0.4}$	56.6	28.7

Table 1: **Reasoning data scaling and selection for the Llama-3.1-8b-Instruct student model.** We compare reasoning data selection based on quality, diversity and difficulty. We use DeepSeek R1 as the teacher model, and conduct supervised finetuning on Llama-3.1-8B-Instruct as the student model. Selection based on diversity in reasoning strategies and question difficulty (via model disagreement) outperforms random selection, although the gap becomes smaller when scaling up data size.

4.1 Reasoning Data Selection

Table 1 and Table 4 show the overall results of the different curation strategies for reasoning distillation from the teacher model, using Llama and Qwen student models respectively. We describe our findings in more details in the following paragraphs.

Baseline Comparisons First we obtain similar observations as in LIMO (Ye et al., 2025) and S1K (Muennighoff et al., 2025): with carefully selected training samples and a strong teacher model, we can effectively distill strong reasoning capabilities even at small data scales. As shown in Table 1, using only 1,000 *randomly selected* examples from NATURALTHOUGHTS already outperforms

LIMO and is on par with S1K, even though these two prior datasets went through rigorous manual selection of questions and reasoning traces.

Diversity In Table 1, we compare selecting examples based on diversity in *questions* (both topics and semantic embeddings), which is often used in prior work (Muennighoff et al., 2025), as well as our proposed diversity in *reasoning strategies*. First, we find that simply uniformly sampling across question topics reduces the performance compared to randomly sampling from the seed set. This is likely due to question topic distribution being too concentrated among a small set of topics. Compared to question topic diversity, the question semantic embedding diversity method produces better performance as the clusters are more fine-grained. Notably, we find that selecting samples based on the diverse *reasoning strategies* results in the best performance out of the three selection criteria. This implies that the diversity of the reasoning traces is more important than the diversity of the questions themselves.

Difficulty We report multiple findings based on training on subsets of varying difficulty. First, the "Long" subset, or examples where shorter reasoning traces are downsampled, performs better than random selection which contains more short reasoning traces. Second, the "Models Disagree" subset performs better than both random selection and "Long". This implies that training on difficult questions that require more advanced reasoning traces is more sample-efficient in distilling reasoning capabilities. Overall, we find that the "Models Disagree" subset leads to the best average performance across all 10K filtered subsets.

Comparing the effects of reasoning verbosity, we find that the "Medium Verbosity" subset is on average the best performing verbosity training set. We hypothesize that the high verbosity subset has too many incoherent reasoning traces, and the low verbosity subset does not contain enough useful reasoning. However, the verbosity distribution is quite concentrated, so it may not be a strong signal to distinguish high-quality examples.

4.2 Reasoning Data Scaling

Different from what was observed in LIMO and S1K, where increasing the number of training samples results in marginal improvement or even decrease in distillation performance, we observe that with NATURALTHOUGHTS, simply scaling up data size beyond 1,000 to 10,000 improves performance on tasks requiring both knowledge and reasoning. This holds for both Llama models, as is shown in Table 1 (*scale*) and Qwen models, as is shown in Table 4, even with random selection. Additionally, using manually curated data as question seeds (NT-NN-LIMO, and NT-NN-S1K) improves the scaling trend compared to random selection. We find that performance across the diverse set of evaluation tasks does not saturate when scaling up the training set size to 500,000 examples, with both random selection and selection based on reasoning strategy diversity (see Figure 1).

We also verify whether the scaling trends persist with a larger student model, which is often more sample efficient. To this end, we use the Llama-3.3-70B-Instruct as the student model. We randomly select 1,000, 10,000 and 100,000 questions from NATURALTHOUGHTS. Results are summarized in Table 2. Similar to the trend observed with Llama-3.1-8B-Instruct and Qwen-2.5-7B-Instruct, scaling up data size beyond 1,000 examples consistently improves performance across all four tasks. With 100k training samples, the student model already outperforms DeepSeek-R1-Distill-Llama-70B on GQPA-D, MMLU-Pro and SuperGPQA.

	Train Size	GPQA-D	MATH-500	MMLU-Pro	SuperGPQA
NATURALTHOUGHTS					
Random Selection	1k	$63.5_{\pm 0.6}$	$86.3_{\pm 0.2}$	78.0	46.0
Random Selection	10k	$65.6_{\pm 0.6}$	$87.4_{\pm 0.3}$	78.4	48.3
Random Selection	100k	$67.6_{\pm 0.4}$	$88.5_{\pm 0.2}$	78.9	50.6
DeepSeek-R1-Distill-Llama-70B	800k	65.2	94.5	78.5	49.4

Table 2: **Scaling results with a larger model.** On general STEM reasoning tasks, training Llama3.3-70B-Instruct with NATURALTHOUGHTS outperforms DeepSeek-R1-Distill-Llama-70B, which was trained with the same teacher and student models using more data.

	No-Think		Adaptive-Think		Think	
	Mean Length	Accuracy	Mean Length	Accuracy	Mean Length	Accuracy
System-1	321.3	$34.0_{\pm 0.6}$	541.1	$32.4_{\pm 0.7}$	-	-
System-2	5133.7	$36.3_{\pm 0.5}$	7817	$37.3_{\pm 0.6}$	8740.6	$37.6_{\pm 0.6}$
Random Mixing						
$p_{System2} = 0.2$	511.0	$31.2_{\pm 0.7}$	1056.0	$31.8_{\pm 0.6}$	8642.3	$34.7_{\pm 0.7}$
$p_{System2} = 0.4$	847.3	$31.8_{\pm 0.5}$	1642.7	$33.2_{\pm 0.5}$	9155.0	$36.7_{\pm 0.6}$
$p_{System2} = 0.6$	729.6	$34.1_{\pm 0.6}$	1683.4	$35.3_{\pm 0.6}$	7139.8	$37.5_{\pm 0.5}$
Difficulty-based Mixing	799.4	$34.5_{\pm 0.5}$	2033.0	$35.1_{\pm 0.6}$	7562.4	$38.9_{\pm 0.7}$

Table 3: **Mixed Reasoning Distillation.** We show the accuracy-efficiency tradeoff of training with mixed System-1 and System-2 reasoning, with three modes at inference time: "No-Think", "Adaptive-Think", and "Think". We report pass@1 accuracy on GPQA-Diamond. The mixed training approach enables operation in all three inference modes, which was not possible with either System-1 or System-2 alone. Our results demonstrate that difficulty-based mixing not only achieves a favorable accuracy-efficiency tradeoff, but also improves overall accuracy in "Think" mode compared to only using System-2. See Figure 6 for a visualization of the trade-offs.

4.3 Mixed Distillation Efficiency

We compare different reasoning distillation approaches outlined in Section 2.3, and evaluate on GPQA-Diamond. We summarize results in Table 3.

System-2 Distillation The baseline System-2 training (i.e. leveraging both the CoT and the final answer) achieves 37.6% accuracy at the cost of efficiency, producing an average response length of 8,740 tokens. In an attempt to elicit responses with shorter reasoning, this model could not provide short answers when instructed not to think ("No-Think" mode), only reducing the average response length to 5,134 tokens.

System-1 Distillation When training with System-1 demonstrations from the teacher model (i.e. only the final answers without reasoning), the student model achieves significant inference-time efficiency gains. We observe that responses are 27x shorter, at the cost of only a 4.6% drop in accuracy compared to System-2 Distillation. However, we find that these models lack the ability to leverage test-time compute, even when explicitly instructed to spend more token budget for reasoning in the Adaptive-Think mode.

Random Mixing Distillation Training with a random mix of System-1 and System-2 reasoning enables the student model to achieve flexible test-time compute by interpolating between thinking fast and slow. For example, when training with 40% System-2 responses ($P_{System2} = 0.4$), the average response length drops from 9,155 tokens in the "Think" mode to 1,642 and 847 for the "Adaptive-Think" and "No-Think" modes, respectively. Increasing the ratio of System-2 examples improves performance but increases response lengths. For example, models trained with 60% System-2 achieve an accuracy of 37.5% in the "Think" mode, compared to 36.7% and 34.7% for training with 40% and 20% System-2 reasoning respectively.

Difficulty Based Mixing Distillation In contrast to random mixing, the difficulty-based mixing approach selectively distills System-2 responses for difficult questions and System-1 responses for easy questions. This targeted strategy, like random mixing, enables the model to flexibly adapt its response length at inference time. Notably, the difficulty-based mixing method achieves an accuracy of 38.9%, representing a 1.3% improvement over System-2 distillation. Furthermore, this approach results in only 36% System-2 responses in the training set, yet it surpasses all random mixing methods, including the one with 60% System-2 responses, which yields an accuracy of 37.5%. This suggests that by selectively applying System-1 and System-2 distillation based on question difficulty, the model can strike a better balance between accuracy and efficiency.

5 Results

We sample questions from NaturalReasoning (Yuan et al., 2025), given its diversity and effectiveness in eliciting reasoning. We use DeepSeek-R1 (Guo et al., 2025) as the teacher model to generate reasoning responses. We use the resulting dataset of distilled reasoning examples, NATURALTHOUGHTS to study data efficiency in distillation. Detailed experimental setup is described in Section 3.

Reasoning Data Scaling. Different from what was observed in LIMO and S1K, where increasing the number of training samples results in marginal improvement or even decrease in distillation performance, we observe that with NATURALTHOUGHTS, simply scaling up data size beyond 1k to 10k improves performance on tasks requiring both knowledge and reasoning. This holds for both Llama models (Table 1-*scale*) and Qwen models (Table 4), even with random selection. We find that performance across the diverse set of evaluation tasks does not saturate when scaling up the training set size to 500k examples, with both random selection and selection based on reasoning strategy diversity (see Figure 1).

Reasoning Data Selection. We compare different curation strategies for reasoning distillation on Llama-3.1-8B-Instr (Table 1) and Qwen-2.5-7B-Instr (Table 4). We also verify whether the scaling trends persist with a larger student model such as Llama-3.3-70B-Instr (Table 2). We summarize our findings as follows with more details in the Appendix.

Baseline Comparisons First we obtain similar observations as in LIMO (Ye et al., 2025) and S1K (Muennighoff et al., 2025): with carefully selected training samples and a strong teacher model, we can effectively distill strong reasoning capabilities even at small data scales. As shown in Table 1, using only 1k *randomly selected* examples from NATURALTHOUGHTS already outperforms LIMO and is on par with S1K, even though these two prior datasets went through rigorous manual selection of questions and reasoning traces.

Diversity In Table 1, we compare selecting examples based on diversity in *questions* (topics and embeddings), which is often used in prior work (Muennighoff et al., 2025), as well as our proposed diversity in *reasoning strategies*. We find that simply uniformly sampling across question topics reduces the performance compared to randomly sampling from the seed set. Compared to question topic diversity, the question semantic embedding diversity method produces better performance as the clusters are more fine-grained. Notably, we find that selecting samples based on *diverse reasoning strategies* results in the best performance out of the three selection criteria.

Difficulty We report several findings based on training on subsets of varying difficulty. First, the "Long" subset, or examples where shorter reasoning traces are downsampled, performs better than random selection. Second, the "Models Disagree" subset, where we use the disagreement between a strong and weak model as a proxy for difficulty, performs better than both random selection and "Long". This implies that training on difficult questions that require more advanced reasoning traces is more sample-efficient in distilling reasoning capabilities (Figure 7). Finally, for reasoning verbosity, where we use LLM-annotated verbosity classifications of reasoning traces as a proxy for difficulty, we find that the "Medium Verbosity" subset is on average the best performing verbosity training set. We hypothesize that the high verbosity subset has too many incoherent reasoning traces, and the low verbosity subset doesn't contain enough useful reasoning. However, the verbosity distribution is concentrated, so it may not be a strong signal to distinguish high-quality examples.

Mixed System-1 and System-2 Distillation. The baseline System-2 training (i.e. leveraging both CoT and final answer) achieves 37.6% accuracy at the cost of efficiency, producing an average response length of 8,740 tokens. In an attempt to elicit shorter reasoning responses, the model could not provide short answers when instructed not to think ("No-Think"), reducing the mean response length to 5,134 tokens. When training with System-1 demonstrations from the teacher model (i.e. only final answers), the student model achieves significant inference-time efficiency gains. We observe that responses are 27x shorter, at the cost of only a 4.6% drop in accuracy compared to System-2 Distillation. However, we find that these models lack the ability to leverage test-time compute, even when explicitly instructed to spend more token budget for reasoning in the Adaptive-Think mode.

Random Mixing Distillation Training with a random mix of System-1 and System-2 reasoning enables the student model to achieve flexible test-time compute by interpolating between thinking fast and slow. For example, when training with 40% System-2 responses ($P_{System2} = 0.4$), the average response length drops from 9,155 tokens in the "Think" mode to 1,642 and 847 for the "Adaptive-Think" and "No-Think" modes, respectively. Increasing the ratio of System-2 examples improves performance but increases response lengths.

Difficulty Based Mixing Distillation In contrast to random mixing, the difficulty-based mixing approach selectively distills System-2 responses for difficult questions and System-1 responses for easy questions. This targeted strategy, like random mixing, enables the model to flexibly adapt its response length at inference time. Notably, the difficulty-based mixing method achieves an accuracy

of 38.9%, representing a 1.3% improvement over System-2 distillation. Furthermore, this approach results in only 36% System-2 responses in the training set, yet it surpasses all random mixing methods, including the one with 60% System-2 responses, which yields an accuracy of 37.5%. This suggests that by selectively applying System-1 and System-2 distillation based on question difficulty, the model can strike a better balance between accuracy and efficiency.

6 Ablations and Analysis

Long is More? Prior work on data selection for LLM alignment has shown that selecting long samples is a tough-to-beat baseline (Zhao et al., 2024). Through our experiments, we validate that length is also a tough-to-beat baseline for LLM reasoning. However, as is shown in Figure 7, we find that the most performant selection strategies (Long, Reasoning Strategies and Models Disagree) all have different distributions of reasoning length, indicating that length is not the only causal factor. Furthermore, selection based on reasoning strategy diversity yields a very similar length distribution to that of random selection, yet the former leads to better performance.

Reducing Sequence Length: Improvements with Mixed Distillation One natural question is whether training with shorter reasoning traces achieves the same efficiency gains as mixed distillation. To this end, we experiment with reducing the maximum sequence length of training examples from 16,384 to 8,192 tokens, and use the same evaluation setting as described in Section 4.3. Results are provided in Table 7.

As expected, this leads to shorter responses at inference time, with an average length of 440 tokens in the Adaptive-Think mode and 372 tokens in No-Think mode. However, it comes at the cost of a severe drop in accuracy, with System-2 distillation scoring 32.5% and 32.1% in those two modes respectively. In contrast, applying difficulty-based mixing with a sequence length of 8,192 tokens not only preserves the ability to adjust response length via Adaptive Thinking but also yields notable accuracy improvements. Specifically, we observe gains of 0.6% over System-2 distillation with a sequence length of 16,384 and 1.2% over System-2 distillation with a sequence length of 8,192, while producing 2x fewer tokens. This suggests that difficulty-based mixing can effectively balance efficiency and accuracy, even at reduced sequence lengths.

Effect of easy-to-verify data We further evaluate the impact of training on problems with easy-to-verify answers. We use selection based on "Models Disagree" as a baseline, to compare with additionally selecting problems whose reference answers have short lengths (≤ 9 words) as a proxy for "easy-to-verify". As is shown in Table 5, training on reasoning traces of such problems does not bring consistent improvements except on math reasoning tasks.

7 Discussion and Conclusion

In this work, we conduct an in-depth analysis of aspects of training data that lead to better scaling performance and sample efficiency when distilling reasoning capabilities from a strong teacher to a weaker student model. Our results offer new insights on the selection and curation of reasoning data, where current understanding mostly focuses on the overall importance of data quality, e.g. the "Less is More" hypothesis. Our experiments indicate that for "learning to reason", the diversity of reasoning primitives matters more than the diversity of topics or domains. Questions that are more difficult usually elicit more reasoning steps and thus serve as better demonstration examples to distill reasoning skills. We attribute lagging improvements on the mathematical task evaluations to the fact that the majority of the problems in our source prompt data do not look like problems appearing in those math evaluations.

Insights derived from our experiments and analysis are highly relevant to the growing interest in building small reasoning models. The training recipe of state-of-the-art small reasoning models (Yang et al., 2025) also emphasizes the importance of data curation used in both RL post-training and strong-to-weak SFT distillation (Bercovich et al., 2025). In practice, a few criteria are used to select samples that (i) are as challenging as possible and (ii) cover a broad range of sub-domains. Our findings are complementary to those, where we demonstrate the potential performance gains from having explicit control over which fine-grained reasoning primitives to distill from the teacher model. Furthermore, we demonstrate steerability of reasoning efficiency in our mixed System-1/System-2 distillation experiments.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes, 2023.

Pranjal Aggarwal and Sean Welleck. L1: Controlling how long a reasoning model thinks with reinforcement learning, 2025. https://arxiv.org/abs/2503.04697.

Can Balioglu, Martin Gleize, Artyom Kozhevnikov, Ilia Kulikov, Tuan Tran, and Julien Yao. fairseq2, 2023. http://github.com/facebookresearch/fairseq2.

Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, Ido Shahaf, Oren Tropp, Ehud Karpas, Ran Zilberstein, Jiaqi Zeng, Soumye Singhal, Alexander Bukharin, Yian Zhang, Tugrul Konuk, Gerald Shen, Ameya Sunil Mahabaleshwarkar, Bilal Kartal, Yoshi Suhara, Olivier Delalleau, Zijia Chen, Zhilin Wang, David Mosallanezhad, Adi Renduchintala, Haifeng Qian, Dima Rekesh, Fei Jia, Somshubra Majumdar, Vahid Noroozi, Wasi Uddin Ahmad, Sean Narenthiran, Aleksander Ficek, Mehrzad Samadi, Jocelyn Huang, Siddhartha Jain, Igor Gitman, Ivan Moshkov, Wei Du, Shubham Toshniwal, George Armstrong, Branislav Kisacanin, Matvei Novikov, Daria Gitman, Evelina Bakhturina, Jane Polak Scowcroft, John Kamalu, Dan Su, Kezhi Kong, Markus Kliegl, Rabeeh Karimi, Ying Lin, Sanjeev Satheesh, Jupinder Parmar, Pritam Gundecha, Brandon Norick, Joseph Jennings, Shrimai Prabhumoye, Syeda Nahida Akter, Mostofa Patwary, Abhinav Khattar, Deepak Narayanan, Roger Waleffe, Jimmy Zhang, Bor-Yiing Su, Guyue Huang, Terry Kong, Parth Chadha, Sahil Jain, Christine Harvey, Elad Segal, Jining Huang, Sergey Kashirsky, Robert McQueen, Izzy Putterman, George Lam, Arun Venkatesan, Sherry Wu, Vinh Nguyen, Manoj Kilaru, Andrew Wang, Anna Warno, Abhilash Somasamudramath, Sandip Bhaskar, Maka Dong, Nave Assaf, Shahar Mor, Omer Ullman Argov, Scot Junkin, Oleksandr Romanenko, Pedro Larroy, Monika Katariya, Marco Rovinelli, Viji Balas, Nicholas Edelman, Anahita Bhiwandiwalla, Muthu Subramaniam, Smita Ithape, Karthik Ramamoorthy, Yuting Wu, Suguna Varshini Velury, Omri Almog, Joyjit Daw, Denys Fridman, Erick Galinkin, Michael Evans, Katherine Luna, Leon Derczynski, Nikki Pope, Eileen Long, Seth Schneider, Guillermo Siman, Tomasz Grzegorzek, Pablo Ribalta, Monika Katariya, Joey Conway, Trisha Saar, Ann Guan, Krzysztof Pawelec, Shyamala Prayaga, Oleksii Kuchaiev, Boris Ginsburg, Oluwatobi Olabiyi, Kari Briski, Jonathan Cohen, Bryan Catanzaro, Jonah Alben, Yonatan Geifman, Eric Chung, and Chris Alexiuk. Llama-nemotron: Efficient reasoning models, 2025. https://arxiv.org/abs/2505.00949.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Do not think that much for 2+3=? on the overthinking of o1-like llms, 2025. https://arxiv.org/abs/2412.21187.

Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, et al. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *arXiv preprint arXiv:2502.14739*, 2025.

Hugging Face. Open r1: A fully open reproduction of deepseek-r1, January 2025. https://github.com/huggingface/open-r1.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings, 2022. https://arxiv.org/abs/2104.08821.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel

Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang,

Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. `https://arxiv.org/abs/2407.21783`. Llama 3.1 Community License Agreement.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, et al. Openthoughts: Data recipes for reasoning models. *arXiv preprint arXiv:2506.04178*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

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

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

Abhinav Kumar, Jaechul Roh, Ali Naseh, Marzena Karpinska, Mohit Iyyer, Amir Houmansadr, and Eugene Bagdasarian. Overthink: Slowdown attacks on reasoning llms. *arXiv e-prints*, pages arXiv–2502, 2025.

Ximing Lu, Seungju Han, David Acuna, Hyunwoo Kim, Jaehun Jung, Shrimai Prabhumoye, Niklas Muennighoff, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, and Yejin Choi. Retrosearch: Exploring untaken paths for deeper and efficient reasoning, 2025. `https://arxiv.org/abs/2504.04383`.

Wenjie Ma, Jingxuan He, Charlie Snell, Tyler Griggs, Sewon Min, and Matei Zaharia. Reasoning models can be effective without thinking, 2025. `https://arxiv.org/abs/2504.09858`.

Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11):205, 2017.

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

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a

benchmark. In *First Conference on Language Modeling*, 2024. https://openreview.net/forum?id=Ti67584b98.

Yiyou Sun, Georgia Zhou, Hao Wang, Dacheng Li, Nouha Dziri, and Dawn Song. Climbing the ladder of reasoning: What llms can-and still can't-solve after sft?, 2025. https://arxiv.org/abs/2504.11741.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, 2020. https://arxiv.org/abs/2002.10957.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024.

Yue Wang, Qiuzhi Liu, Jiahao Xu, Tian Liang, Xingyu Chen, Zhiwei He, Linfeng Song, Dian Yu, Juntao Li, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Thoughts are all over the place: On the underthinking of o1-like llms, 2025. https://arxiv.org/abs/2501.18585.

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, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. https://arxiv.org/abs/2505.09388.

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

Ping Yu, Jing Xu, Jason Weston, and Ilia Kulikov. Distilling system 2 into system 1. *arXiv preprint arXiv:2407.06023*, 2024a.

Ping Yu, Jing Xu, Jason Weston, and Ilia Kulikov. Distilling system 2 into system 1, 2024b. https://arxiv.org/abs/2407.06023.

Weizhe Yuan, Jane Yu, Song Jiang, Karthik Padthe, Yang Li, Dong Wang, Ilia Kulikov, Kyunghyun Cho, Yuandong Tian, Jason E Weston, et al. Naturalreasoning: Reasoning in the wild with 2.8 m challenging questions. *arXiv preprint arXiv:2502.13124*, 2025. Creative Commons Attribution Non Commercial 4.0 license.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model?, 2025. https://arxiv.org/abs/2504.13837.

Hao Zhao, Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. Long is more for alignment: A simple but tough-to-beat baseline for instruction fine-tuning, 2024. https://arxiv.org/abs/2402.04833.

A Related Work

Data-centric Approaches for Improving Reasoning Distillation and reinforcement learning have become standard approaches for building strong reasoning models. Guo et al. (2025) explicitly compare the effectiveness of distillation vs. RL for improving small models' reasoning capabilities, and find that distilling from a strong teacher model significantly outperforms large-scale RL training only on the student model. Since then, the community has been actively working on distillation and proposing distilled datasets such as OpenThoughts (Guha et al., 2025) and OpenR1 (Face, 2025). However, there has been little study on the quality of distillation datasets until recently. Prior work such as LIMO (Ye et al., 2025) and S1K (Muennighoff et al., 2025) shows that carefully curated questions and reasoning traces can greatly improve sample efficiency, where 1,000 examples are sufficient to distill the long CoT behaviors from the teacher model. However, these hypotheses were primarily verified in narrow domains such as math and coding. Our work provides new insights on a much broader set of domains with more diverse reasoning problems.

Reasoning Efficiency Recent work has shown that overthinking is a common pattern in SoTA reasoning models (e.g., R1 (Guo et al., 2025), o1 (Jaech et al., 2024), o3, etc.) and thus the need to curate high-quality reasoning traces (Kumar et al., 2025; Chen et al., 2025) as well as to develop methods that elicit more efficient reasoning (Aggarwal and Welleck, 2025). In particular, Ma et al. (2025) demonstrate non-trivial reasoning performance by simply omitting thinking in the training data. An alternative approach to selecting reasoning traces is to revise the teacher model's reasoning traces and generate new reasoning paths (Lu et al., 2025). Distilling System 2 into System 1 was first proposed in Yu et al. (2024b).

B Additional Tables and Figures

	Train Size	GPQA-D	MATH-500	MMLU-Pro	SuperGPQA
Baselines					
Qwen-2.5-7B-Instruct	0	$34.1_{\pm 0.5}$	$69.7_{\pm 0.3}$	56.5	29.5
OpenThoughts3 (Guha et al., 2025)	10k	$34.7_{\pm 0.5}$	$83.2_{\pm 0.2}$	59.1	28.5
OpenThoughts3 (Guha et al., 2025)	100k	$41.8_{\pm 0.5}$	$88.1_{\pm 0.2}$	62.8	33.7
OpenThoughts3 (Guha et al., 2025)	1.2m	$46.9_{\pm 0.5}$	$91.2_{\pm 0.1}$	59.1	33.5
DeepSeek-R1-Distill-Qwen-7B	800k	49.1	92.8	56.8	29.8
NATURALTHOUGHTS					
Random	10k	$41.8_{\pm 0.6}$	$79.0_{\pm 0.2}$	59.7	31.2
Reasoning Strategies	10k	$42.2_{\pm 0.6}$	$78.6_{\pm 0.2}$	61.5	32.1
Models Disagree	10k	$40.5_{\pm 0.5}$	$78.4_{\pm 0.3}$	61.7	32.7
Long	10k	$40.0_{\pm 0.6}$	$78.8_{\pm 0.2}$	61.8	33.5
Random	100k	$42.6_{\pm 0.5}$	$80.8_{\pm 0.2}$	62.7	32.8
Models Disagree	100k	$43.7_{\pm 0.6}$	$80.8_{\pm 0.2}$	62.8	33.9
Reasoning Strategies	100k	$43.5_{\pm 0.5}$	$80.9_{\pm 0.2}$	62.2	33.1
Long	100k	$43.4_{\pm 0.4}$	$81.7_{\pm 0.4}$	62.2	34.0
Random	500k	$48.6_{\pm 0.6}$	$83.1_{\pm 0.2}$	62.3	35.2
Reasoning Strategies	500k	$48.3_{\pm 0.5}$	$83.6_{\pm 0.2}$	62.7	35.2

Table 4: **Reasoning data scaling and selection for the Qwen-2.5-7B-Instruct student model.** We also provide comparisons when using Qwen-2.5-7B-Instruct as the student model. Training with 500k examples from NATURALTHOUGHTS outperforms training with 1.2m examples from OpenThoughts3 (Guha et al., 2025) on three of the four evaluation benchmarks.

	Train Size	GPQA-D	MATH-500	MMLU Pro	SuperGPQA
Models Disagree	10k	$39.7_{\pm 0.4}$	$61.9_{\pm 0.3}$	59.1	29.7
with short reference answer	10k	$39.0_{\pm 0.7}$	$63.0_{\pm 0.3}$	59.2	28.1
Models Disagree	100k	$45.2_{\pm 0.6}$	$70.2_{\pm 0.2}$	59.8	32.2
with short reference answer	100k	$43.8_{\pm 0.5}$	$71.4_{\pm 0.4}$	61.4	31.2

Table 5: **Reference-answer-based selection.** We show that filtering questions based on the length of the reference answers (as a proxy for easy-to-verify problems), in addition to disagreement selection, yields improvement in math reasoning performance.

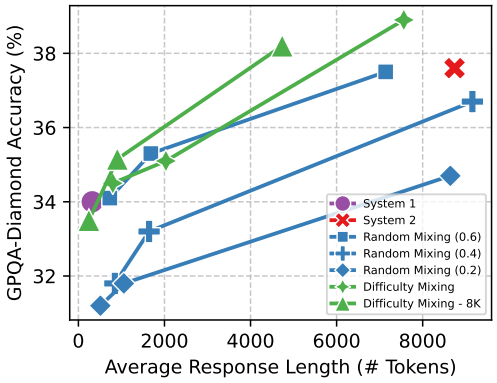

Figure 3: A mix of System-1 (no reasoning traces, final answer only) and System-2 (full reasoning traces) distillation improves the inference-time efficiency of the student model (details in Section 2.3 and Section 4.3).

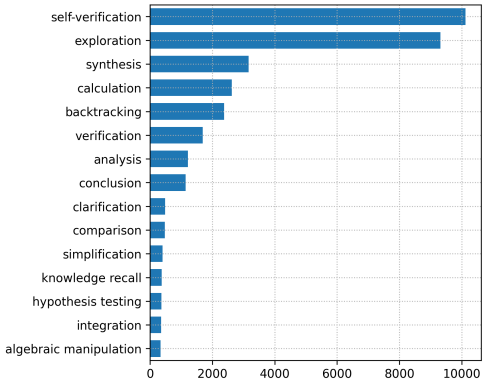

Figure 4: Top 20 commonly used reasoning strategies based on annotations of 10,000 samples. We observe that there are a small number frequently used strategies, followed by a long tail of more niche strategies.

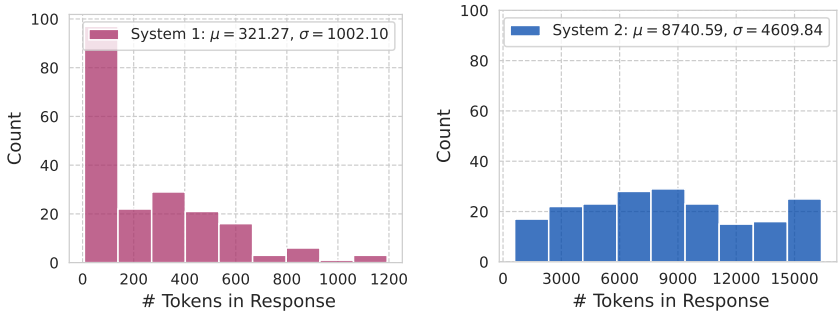

Figure 5: **Generation length distributions of System-1 and System-2 reasoning for GPQA-Diamond.** System-1 responses (left) are predominantly brief, indicating minimal thinking, if any. In contrast, System-2 responses (right) are significantly longer than those of System-1 and exhibit greater variance, as the response lengths vary based on the amount of thinking required for each question, depending on its complexity.

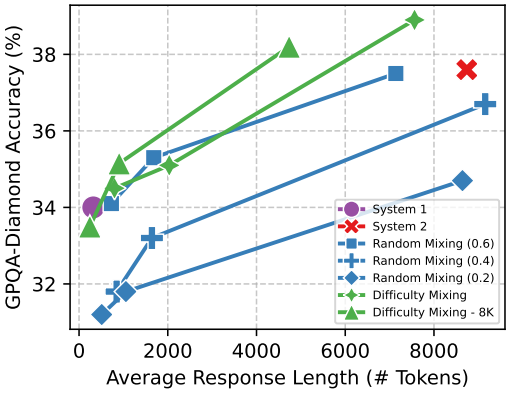

Figure 6: Mixed System-1 (final answer only) and System-2 (reasoning trace + final answer) distillation improves the inference-time efficiency of the student model (details in Table 3 and Section 4.3).

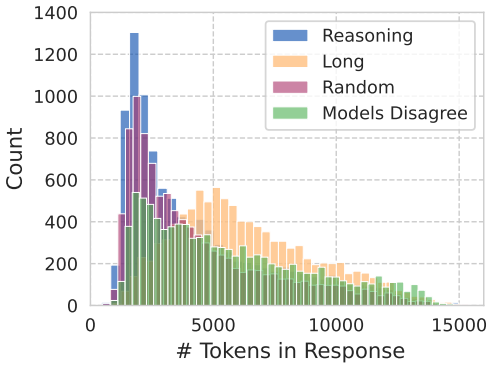

Figure 7: Length distributions of data subsets obtained from different selection methods.

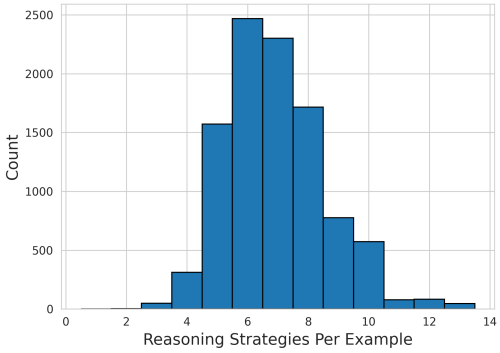

Figure 8: Distribution of unique reasoning strategies per example based on annotations of 10,000 samples.

C Prompts used in Annotation

The prompt used for annotating reasoning traces to obtain reasoning strategies and verbosity score is provided in Figure 9. Figure 10 covers the prompt used to annotate disciplines and fields for questions in the dataset.

```
Below is a question and solution generated by an LLM. Your task is
to summarize the problem-solving steps used by the LLM. Read the
thought process carefully, and annotate the explorations in the
thought process used by the LLM. Specifically, write down detailed
steps the LLM took to pursue its thinking process, identifying all
meta-reasoning strategies used at each step, e.g. self-verification
, backtracking, exploration, etc. Based on these analysis, also
check the degrees of verbosity of the reasoning traces, e.g. how
much unnecessary ramblings were found during the thinking process
which does not make much progress in solving the problem. Derive a
verbosity_score in the end. The verbosity_score should be derived
on a scale of 0 to 10. Score 0 means the problem solving in the
thinking process is very efficient with no rambling at all. Score
10 means the reasoning traces are very verbose, where the thinking
process is long but each step does not make progress in solving the
 problem. Organize your answer in a json so that the steps and meta
-reasoning strategies and the final verbosity_score can be easily
extracted.
Question: {question}
Solution from LLM: {reasoning trace to be annotated}}
```

Figure 9: Prompt for annotating reasoning traces to obtain reasoning strategies and the verbosity score.

```
You are an expert in labeling questions into categories.

For a given question, use the following taxonomy for labelling,
which is structured as {'discipline': {'field': ['sub-field', 'sub-
field', ...]}}.

{'Engineering': {'Electronic Science and Technology': ['Circuits
and Systems',
   'Microelectronics and Solid-State Electronics',
   'Electromagnetic Field and Microwave Technology'],
  'Computer Science and Technology': ['Computer Architecture',
   'Data Structures',
   'Operating Systems',
   'Computer Software and Theory',
   'Advanced Programming Languages',
   'Pattern Recognition',
   'Principles of Computer Organization',
   'Computer Networks',
   'Databases',
   'Formal Languages'],
  'Information and Communication Engineering': ['Signal and
  Information Processing',
   'Optical Fiber Communication',
   'Communication and Information Systems',
   'Antenna and Radio Communication',
   'Communication Principles'],
  'Control Science and Engineering': ['Control Theory and Control
  Engineering',
   'Operations Research and Cybernetics',
   'Guidance, Navigation and Control'],
  'Materials Science and Engineering': ['Materials Physics and
  Chemistry',
   'Materials Processing Engineering'],
  'Electrical Engineering': ['Power Electronics and Electrical
  Drives',
   'Electrical Theory and New Technologies',
   'High Voltage and Insulation Technology',
   'Power Systems and Automation'],
  'Power Engineering and Engineering Thermophysics': ['Power
  Machinery and Engineering',
   'Internal Combustion Engineering',
   'Thermal Energy Engineering',
   'Engineering Thermophysics',
   'Refrigeration and Cryogenic Engineering',
   'Fluid Machinery and Engineering',
   'Heat Transfer',
   'Engineering Fluid Mechanics'],
  'Hydraulic Engineering': ['Water conservancy and Hydropower
  Engineering',
   'Hydraulics and Hydrology'],
  'Chemical Engineering and Technology': ['Mass Transport and
  Separation Process in Chemical Engineering',
   'Fluid Flow and Heat Transfer in Chemical Engineering',
   'Elements of Chemical Reaction Engineering',
   'Chemical Transport Engineering'],
  'Architecture': ['Urban Planning and Design',
   'Architectural Design and Theory',
   'Architectural History'],
  'Forestry Engineering': ['Forest Engineering',
   'Wood Science and Technology'],
```

```
'Nuclear Science and Technology': ['Radiation Protection and
    Nuclear Technology Applications',
 'Nuclear Energy and Reactor Technology'],
'Weapon Science and Technology': ['Weapon Systems Science and
    Engineering',
 'Military Chemistry and Pyrotechnics'],
'Naval Architecture and Ocean Engineering': ['Marine Engineering
    ',
 'Ship Mechanics and Design Principles'],
'Environmental Science and Engineering': ['Environmental and
    Resource Protection',
 'Environmental Engineering',
 'Environmental Science'],
'Transportation Engineering': ['Vehicle Operation Engineering',
 'Traffic Information Engineering and Control',
 'Transportation Planning and Management',
 'Road and Railway Engineering'],
'Mechanical Engineering': ['Manufacturing Automation',
 'Mechatronic Engineering'],
'Aeronautical and Astronautical Science and Technology': ['
    Aeronautical and Astronautical Science and Technology'],
'Civil Engineering': ['Geotechnical Engineering',
 'Structural Engineering',
 'Bridge and Tunnel Engineering',
 'Urban Infrastructure Engineering'],
'Mechanics': ['Fundamentals of Dynamics and Control',
 'Theoretical Fluid Mechanics',
 'Theoretical Mechanics',
 'Rigid Body Mechanics',
 'Solid Mechanics'],
'Petroleum and Natural Gas Engineering': ['Poromechanics and
    Reservoir Physics',
 'Oil and Gas Field Development and Storage & Transportation
     Engineering'],
'Food Science and Engineering': ['Food Processing and Storage
    Engineering',
 'Food Biochemistry'],
'Agricultural Engineering': ['Agricultural Environment and Soil-
    Water Engineering',
 'Agricultural Mechanization Engineering'],
'Surveying and Mapping Science and Technology': ['Geodesy and
    Surveying Engineering',
 'Cartography and Geographic Information Engineering',
 'Digital Surveying and Remote Sensing Applications'],
'Metallurgical Engineering': ['Iron and Steel Metallurgy',
 'Principles of Metallurgy',
 'Non-ferrous Metallurgy',
 'Physical Chemistry of Metallurgical Process'],
'Mining Engineering': ['Mining and Safety Engineering',
 'Mineral Processing Engineering'],
'Geological Resources and Geological Engineering': ['Geological
    Resources and Geological Engineering'],
'Optical Engineering': ['Theoretical Optics',
 'Optoelectronic Technology',
 'Laser Technology',
 'Applied Optics'],
'Textile Science and Engineering': ['Textile Materials Science',
 'Textile Chemistry and Dyeing Engineering'],
'Instrument Science and Technology': ['Instrument Science and
    Technology']},
```

```
'Philosophy': {'Philosophy': ['Philosophical Aesthetics',
   'Ethics',
   'Logic',
   'Philosophy of Science and Technology',
   'Religious Studies']},
'Medicine': {'Traditional Chinese Medicine': ['Traditional Chinese
    Medicine Theory',
   'Traditional Chinese Health Preservation',
   'Traditional Chinese Pharmacy'],
  'Clinical Medicine': ['Internal Medicine',
   'Obstetrics and Gynecology',
   'Emergency Medicine',
   'Neurology',
   'Psychiatry and Mental Health',
   'Surgery',
   'Imaging and Nuclear Medicine',
   'Otorhinolaryngology',
   'Dermatology and Venereology',
   'Ophthalmology',
   'Geriatric Medicine',
   'Oncology',
   'Clinical Laboratory Diagnostics',
   'Anesthesiology',
   'Pediatrics',
   'Nursing and Rehabilitation Medicine'],
  'Basic Medicine': ['Pathogen Biology',
   'Immunology',
   'Human Anatomy and Histology-Embryology',
   'Pathology and Pathophysiology',
   'Forensic Medicine',
   'Radiation Medicine'],
  'Public Health and Preventive Medicine': ['Epidemiology and
      Health Statistics',
   'Health Toxicology and Environmental Health',
   'Maternal, Child and Adolescent Health',
   'Nutrition and Food Hygiene'],
  'Pharmacy': ['Pharmacology',
   'Microbiology and Biochemical Pharmacy',
   'Pharmaceutical Analysis',
   'Medicinal Chemistry',
   'Pharmaceutics'],
  'Stomatology': ['Basic Stomatology', 'Clinical Stomatology']},
'Economics': {'Applied Economics': ['Finance',
   'Public Finance',
   'International Trade',
   'Labor Economics',
   'Economic Statistics',
   'Quantitative Economics',
   'Industrial Economics',
   'National and Defense Economics'],
  'Theoretical Economics': ['Political Economy',
   'Economic History',
   'Western Economics']},
```

```
'Science': {'Mathematics': ['Combinatorial Mathematics',
   'Ordinary Differential Equations',
   'Mathematical Analysis',
   'Advanced Algebra',
   'Functions of Real Variables',
   'Probability and Statistics',
   'Numerical Analysis',
   'Polynomials and Series Expansions',
   'Geometry and Topology',
   'Computational Mathematics',
   'Discrete Mathematics',
   'Stochastic Processes',
   'Functions of Complex Variables',
   'Fundamental Mathematics',
   'Number Theory',
   'Group Theory',
   'Cryptography',
   'Fuzzy Mathematics',
   'Special Number Theory',
   'Graph Theory'],
  'Physics': ['Particle and Nuclear Physics',
   'Electrodynamics',
   'Quantum Mechanics',
   'Atomic and Molecular Physics',
   'Fluid Physics',
   'Solid State Physics',
   'Relativity',
   'Thermodynamics and Statistical Physics',
   'Subatomic and Atomic Physics',
   'Semiconductor Physics',
   'Polymer Physics',
   'Statistical Mechanics',
   'Thermodynamics',
   'Acoustics'],
  'Systems Science': ['Systems Science'],
  'Biology': ['Botany',
   'Biochemistry and Molecular Biology',
   'Genetics',
   'Zoology',
   'Biophysics',
   'Cell Biology',
   'Physiology',
   'Microbiology',
   'Ecology'],
  'Chemistry': ['Organic Chemistry',
   'Physical Chemistry',
   'Analytical Chemistry',
   'Electrochemistry',
   'Radiochemistry',
   'Polymer Chemistry and Physics',
   'Inorganic Chemistry'],
  'Geography': ['Human Geography', 'Physical Geography'],
  'Oceanography': ['Hydrogeology',
   'Marine Chemistry',
   'Underwater Acoustics',
   'Marine Biology'],
```

```
  'Geology': ['Principles of Seismic Exploration',
   'Structural Geology',
   'Mineralogy, Petrology, and Economic Geology',
   'Paleontology and Stratigraphy',
   'Geochemistry'],
  'Physical Oceanography': ['Physical Oceanography'],
  'Astronomy': ['Solar System Science',
   'Astrophysics',
   'Stellar and Interstellar Evolution',
   'Astronomical Observation and Technology',
   'Cosmology'],
  'Atmospheric Science': ['Atmospheric Physics and Atmospheric
     Environment',
   'Meteorology',
   'Dynamic Meteorology'],
  'Geophysics': ['Solid Earth Geophysics', 'Space physics']},
'Law': {'Law': ['Procedural Law',
   'Constitutional and Administrative Law',
   'Criminal Law',
   'Civil and Commercial Law',
   'Contract Law',
   'Military Law',
   'Law and Social Governance',
   'International Law',
   'Legal Theory and Legal History'],
  'Political Science': ['Political Science']},
'History': {'History': ['World History',
   'Historical Geography',
   'Archaeology and Museology']},
'Education': {'Education': ['Theory of Curriculum and Instruction',
   'Educational Technology and Principles',
   'Special Education',
   'Preschool Education'],
  'Psychology': ['Psychology'],
  'Physical Education': ['Sports Science and Medicine',
   'Sports Humanities and Sociology',
   'Physical Education and Training']},
'Military Science': {'Military Science': ['Military Thought and
    History',
   'Military Command and Information Systems',
   'Military Management',
   'Military Logistics and Equipment']},
'Management': {'Business Administration': ['Business and Accounting
     Management',
   'Tourism Management and Technological Economics Management'],
  'Public Administration': ['Social Medicine and Health Management
     ',
   'Land Resource Management and Administrative Management',
   'Education Economics, Management and Social Security'],
  'Library, Information and Archival Management': ['Information
     Management and Communication',
   'Library and Archival Science',
   'Information Management Science'],
  'Management Science and Engineering': ['Management Science and
     Engineering']},
```

```
  'Literature and Arts': {'Language and Literature': ['Literary
  History',
   'French Language and Literature',
   'Linguistics and Applied Linguistics',
   'Literary Theory',
   'Philology and Bibliography',
   'Modern and Contemporary Chinese Literature',
   'Classical Chinese Literature',
   'Russian Language and Literature'],
  'Art Studies': ['Dance Studies',
   'Design Arts',
   'Broadcasting and Television Art',
   'Fine Arts',
   'Drama and Opera Studies',
   'Film Studies'],
  'Journalism and Communication': ['Journalism and News Practice',
   'History and Theory of Journalism and Media Management',
   'Communication and Broadcasting'],
  'Musicology': ['Composition',
   'Instrumentation and Performance',
   'Music History, Education, and Technology',
   'Musical Forms and Analysis',
   'Harmony',
   'Pitch and Scales']},
'Agronomy': {'Aquaculture': ['Aquaculture'],
  'Animal Husbandry': ['Animal Rearing and Breeding',
   'Animal Nutrition and Feed Science'],
  'Crop Science': ['Crop Science'],
  'Forestry': ['Forest Cultivation and Genetic Breeding',
   'Landscape Plants and Ornamental Horticulture'],
  'Veterinary Medicine': ['Veterinary Medicine']},
'Sociology': {'Sociology': ['Social and Folklore Studies',
   'Demography and Anthropology']}}
First, identify the most appropriate discipline. Then identify the
most appropriate filed within the chosen discipline. Finally,
identify the most appropriate sub-filed.

Put your final labelling results in a json object:
```
{
   "discipline": <choose the most appropriate discipline>,
   "field": <choose the most appropriate field within the
   discipline>,
    "sub_field": <choose the most appropriate sub_field within the
    field>
}
```

Under no circumstances use enumeration and do not give more than 1
discipline, 1 field and 1 sub-field and do not use any labels that
are not in above dictionary.

### Question
```

Figure 10: Prompt for annotating question with discipline and fields.

Table 6: Comparison of different semantic embedding clustering methods for 10k training examples.

Embeddings	Clustering	# clusters	GPQA-D	MATH-500	MMLU Pro	SuperGPQA	Average
all-MiniLM-L6-v2	K-means	10k	$36.3_{\pm 0.8}$	$57.5_{\pm 0.4}$	55.4	25.9	43.8
sup-simcse-roberta-large	K-means	10k	$34.3_{\pm 0.7}$	$58.8_{\pm 0.4}$	56.1	26.7	44.0
Llama-3.1-8B-Instruct	K-means	100	$39.4_{\pm 0.7}$	$57.6_{\pm 0.4}$	55.5	26.1	44.7
		500	$36.0_{\pm 0.5}$	$59.6_{\pm 0.3}$	56.4	26.5	44.6
		1k	$35.8_{\pm 0.8}$	$59.8_{\pm 0.3}$	56.5	26.7	44.7
		10k	$36.5_{\pm 0.6}$	$59.0_{\pm 0.3}$	57.3	26.2	44.8
	HDBSCAN	991	$39.4_{\pm 0.5}$	$60.3_{\pm 0.3}$	55.9	27.2	45.7

Table 7: Comparison of different methods for mixing System-1 and System-2 responses using sequence lengths of 8K and 16K.

		No-Think		**Adaptive-Think**		**Think**	
	Sequence Length	Mean Length	Accuracy	Mean Length	Accuracy	Mean Length	Accuracy
System-1	8192	250.8	$34_{\pm 0.5}$	533.4	$33.1_{\pm 0.5}$	-	-
System-1	16384	321.3	$34_{\pm 0.6}$	541.1	$32.4_{\pm 0.7}$	-	-
System-2	8192	371.6	$32.1_{\pm 0.5}$	439.5	$32.5_{\pm 0.5}$	8510.2	$37_{\pm 0.6}$
System-2	16384	5133.7	$36.3_{\pm 0.5}$	7817	$37.3_{\pm 0.6}$	8740.6	$37.6_{\pm 0.6}$
Difficulty-based Mixing	8192	243.6	$33.5_{\pm 0.6}$	904.6	$35.15_{\pm 0.5}$	4735.4	$38.2_{\pm 0.5}$
Difficulty-based Mixing	16384	799.4	$34.5_{\pm 0.5}$	2033	$35.1_{\pm 0.6}$	7562.4	$38.9_{\pm 0.7}$

D Diversity using Semantic Embeddings based Clustering

In Table 6, we compare the performance of different clustering methods. Three popular embedding models are considered: all-MiniLM-L6-v2 (Wang et al., 2020), SimCSE (Gao et al., 2022) and Llama-3.1-8B-Instruct (Grattafiori et al., 2024). The embedding size of all-MiniLM-L6-v2 is 384, sup-simcse-roberta-large is 1024 and Llama-3.1-8B-Instruct is 4096. After choosing the best embedding model Llama-3.1-8B, we then experiment with different clustering methods. We use K-means with different numbers of clusters: 100, 500, 1k, 10k. We also experiment with density-based HDBSCAN (McInnes et al., 2017) clustering method. In each experiment, we use equal stratified sampling to randomly sample same number of items from each cluster. The best combination of Llama-3.1-8B-Instruct with HDBSCAN is reported in the main experiment Table 1.

E Reasoning strategy count ablation

The number of unique reasoning strategies found per training example is shown in Figure 8. We split training examples into bucket Low (less than 5 strategies), Med (5-8 strategies) and High (more than 8 strategies) and perform SFT distillation. From Table 8 we observe that on average the Med category performs the best. The fact that performance varies on different buckets indicates that the reasoning strategy should align with specific tasks.

Table 8: Comparison of different reasoning strategy count sampling methods.

sampling method	# examples	GPQA-D	MATH-500	MMLU Pro	SuperGPQA
# Strategies=Low	10k	$34.9_{\pm 0.4}$	$60.9_{\pm 0.2}$	57.2	27.5
# Strategies=Med	10k	$40.3_{\pm 0.6}$	$59.9_{\pm 0.3}$	56.8	28.2
# Strategies=High	10k	$38.2_{\pm 0.5}$	$60.8_{\pm 0.3}$	57.4	27.6

F Limitations

Our experiments are conducted in the setting of off-policy distillation where the student model is trained on the cross entropy loss from the teacher's labels. Another commonly adopted approach

```
{'Mathematics': (1338522, 49.92),
 'Physics': (443010, 16.52),
 'Computer Science and Technology': (196323, 7.32),
 'Applied Economics': (144020, 5.37),
 'Biology': (60535, 2.26),
 'Chemistry': (51152, 1.91),
 'Law': (45345, 1.69),
 'Electronic Science and Technology': (43368, 1.62),
 'Education': (42881, 1.6),
 'Philosophy': (38093, 1.42),
 'Mechanical Engineering': (34756, 1.3),
 'Electrical Engineering': (24703, 0.92),
 'Clinical Medicine': (16956, 0.63),
 'Basic Medicine': (16256, 0.61),
 'Astronomy': (15106, 0.56),
 'Power Engineering and Engineering Thermophysics': (14027, 0.52),
 'Mechanics': (12406, 0.46),
 'Business Administration': (11542, 0.43),
 'Theoretical Economics': (10922, 0.41),
 'Information and Communication Engineering': (9936, 0.37),
 'Aeronautical and Astronautical Science and Technology': (8535,
    0.32),
 'Civil Engineering': (7997, 0.3),
 'Control Science and Engineering': (7701, 0.29),
 'Language and Literature': (7255, 0.27),
 'Public Health and Preventive Medicine': (6363, 0.24),
 'Management Science and Engineering': (6072, 0.23),
 'Psychology': (5744, 0.21),
 'History': (5271, 0.2),
 'Geography': (4998, 0.19),
 'Sociology': (4482, 0.17),
 'Materials Science and Engineering': (4458, 0.17),
 'Optical Engineering': (3987, 0.15),
 'Chemical Engineering and Technology': (3541, 0.13),
 'Geology': (3340, 0.12),
 'Hydraulic Engineering': (3069, 0.11),
 'Transportation Engineering': (2621, 0.1),
 'Musicology': (2178, 0.08),
 'Crop Science': (2176, 0.08),
 'Art Studies': (1966, 0.07),
 'Atmospheric Science': (1955, 0.07),
 'Physical Education': (1939, 0.07),
 'Nuclear Science and Technology': (1836, 0.07),
 'Pharmacy': (1572, 0.06),
 'Geophysics': (1493, 0.06),
 'Environmental Science and Engineering': (1269, 0.05),
 'Systems Science': (1106, 0.04),
 'Political Science': (804, 0.03),
 'Naval Architecture and Ocean Engineering': (727, 0.03),
 'Animal Husbandry': (706, 0.03),
 'Public Administration': (651, 0.02),
 'Military Science': (593, 0.02),
 'Oceanography': (576, 0.02),
 'Architecture': (532, 0.02),
 'Food Science and Engineering': (482, 0.02),
 'Petroleum and Natural Gas Engineering': (441, 0.02),
```

```
'Surveying and Mapping Science and Technology': (405, 0.02),
'Journalism and Communication': (374, 0.01),
'Textile Science and Engineering': (328, 0.01),
'Stomatology': (275, 0.01),
'Metallurgical Engineering': (266, 0.01),
'Forestry': (170, 0.01),
'Weapon Science and Technology': (137, 0.01),
'Veterinary Medicine': (133, 0.0),
'Instrument Science and Technology': (125, 0.0),
'Aquaculture': (125, 0.0),
'Library, Information and Archival Management': (110, 0.0),
'Physical Oceanography': (98, 0.0),
'Traditional Chinese Medicine': (69, 0.0),
'Mining Engineering': (68, 0.0),
'Agricultural Engineering': (63, 0.0),
'Forestry Engineering': (48, 0.0),
'Geological Resources and Geological Engineering': (24, 0.0)}
```

Figure 11: Domain distribution(count and percentage) among all the questions.

in knowledge distillation is on-policy distillation (Agarwal et al., 2023), where the student model is trained to match the logits of the teacher model computed given a partial context generated by the student. Future work should verify whether the same findings hold in the on-policy setting. The implications of selecting distillation examples can be further explored by conducting reinforcement learning following the SFT distillation stage with curated reasoning data.

Question: A weight is attached to a string and spun vertically in a gravity field. Describe the resultant force acting on the weight at the leftmost and rightmost positions of its circular motion, and explain why the resultant force does not always point towards the center of the circle.

"Okay, so I have this problem where a weight is attached to a string and spun vertically in a gravity field. I need to describe the resultant force acting on the weight at the leftmost and rightmost positions of its circular motion. Also, I have to explain why the resultant force doesn't always point towards the center of the circle. Hmm, let me think about this step by step.

First, I remember that when an object moves in a circular path, it experiences a centripetal force directed towards the center of the circle. This centripetal force is necessary to keep the object moving in a circular path instead of a straight line. In many cases, like when spinning a ball horizontally, the tension in the string provides the centripetal force. But in this case, it's a vertical circle, so gravity must play a role here.

At different points in the vertical circle, the forces acting on the weight will be different because gravity is always acting downward, whereas the tension in the string will vary depending on the position. Let me try to visualize this. When the weight is at the top of the circle, both tension and gravity are acting downward. At the bottom, tension acts upward and gravity downward. At the leftmost and rightmost positions, which are the sides of the circle, the string is horizontal, right? So gravity is still acting downward, and tension is along the string towards the center. Wait, but if the weight is at the leftmost point, the center is to the right, so tension is to the right, gravity is down. So the resultant force would be the combination of tension and gravity. Similarly, at the rightmost point, tension is towards the left, and gravity is still down.

But the question is about the leftmost and rightmost positions. Let me focus on those. So in these positions, the weight is moving either upward or downward? Wait, at the leftmost point, if the weight is moving in a vertical circle, the direction of motion is either up or down, depending on whether it's moving clockwise or counterclockwise. Let's assume it's moving in a vertical circle in such a way that when it's at the leftmost point, it's moving downward, and at the rightmost point, it's moving upward. Wait, that might not be right. Let me clarify.

In vertical circular motion, when you spin something around, at the top, the velocity is to the left or right, and at the bottom, it's the opposite. So actually, at the leftmost point, the velocity is downward, and at the rightmost point, it's upward? Hmm, maybe. Alternatively, if we're considering the standard vertical circle where you start from the bottom and go up, then at the top, the velocity is to the left, at the bottom to the right, so the leftmost and rightmost points would have velocity upward and downward? Wait, perhaps it's better to parametrize the motion.

Alternatively, perhaps more straightforwardly, at any point in the circular motion, the direction of the velocity is tangential to the circle. So at the leftmost point, the velocity is straight down, and at the rightmost point, it's straight up. Wait, no. If the circle is vertical, the top of the circle is highest point, bottom is lowest. Then, leftmost and rightmost points are at the sides. So when the object is at the leftmost point, moving down from the top to the bottom, the velocity direction is downward. Wait, actually, at the leftmost point, the velocity is in the downward direction? Wait, no, in a circle, the direction of velocity is tangent to the circle. So at the leftmost point, if you imagine a circle, the tangent is vertical, pointing down on the left and up on the right. Wait, no. Wait, maybe I should draw this mentally.

Consider a circle on a vertical plane. The top point has velocity to the left or right, depending on the direction of rotation. Suppose the weight is moving clockwise. Then at the top, velocity is to the right; at the bottom, to the left; at the leftmost point, velocity is downward; at the rightmost point, velocity is upward. If it's counterclockwise, then at the top, velocity is to the left; bottom to the right; leftmost upward; rightmost downward. So depending on direction, the velocity direction at those side points is either up or down. Hmm.

But regardless, in terms of the forces acting on the weight at the leftmost and rightmost positions: At each of these points, the forces acting are tension in the string and gravity. The tension is directed towards the center of the circle, which at leftmost position is towards the right, and at the rightmost position is towards the left. Gravity is always acting downward, with magnitude mg. So at these two points, the tension provides the centripetal force component, but gravity acts perpendicular to the tension? Wait, at the leftmost and rightmost points, gravity is vertical and tension is horizontal. Therefore, the two forces are perpendicular to each other at these points. Therefore, the resultant force would be the vector sum of the horizontal tension and the vertical gravitational force. So in that case, the resultant force is not purely towards the center, because you have two components: one towards the center (tension) and one downward (gravity). So the resultant force would be the combination of these two, hence not pointing towards the center. But wait, isn't centripetal acceleration still required? Let me think.

Wait, in uniform circular motion, the net force must be centripetal, but in non-uniform circular motion, such as when there is tangential acceleration, the net force has both centripetal and tangential components. In this case, when you spin a weight vertically, the speed is not constant because gravity does work on the object, speeding it up as it comes down and slowing it down as it goes up. Therefore, the motion is not uniform, so there is a tangential component of acceleration, which means the net force has both a centripetal component (towards the center) and a tangential component (tangent to the circle).

Therefore, at the leftmost and rightmost points, the weight is experiencing both a centripetal force (due to tension) and a gravitational force (downward). Since these are perpendicular, the resultant force is the combination of the two. However, wait, at the leftmost and rightmost positions, is there a tangential acceleration? Let's see. If the object is moving in a circle with varying speed, then the tangential acceleration is present. At the leftmost point, if the object is moving downward (in a clockwise rotation), then the gravitational force has a component along the direction of motion, which would cause a tangential acceleration. Wait, but at the exact leftmost point, the velocity is purely downward, and gravity is acting straight down. However, the tension at that point is horizontal. So at that exact point, the gravitational force is entirely tangential (since the tangent at the leftmost point is vertical). Therefore, the gravitational force is entirely contributing to the tangential acceleration, while the tension provides the centripetal force. But wait, if that's the case, then the net force would have two components: tension (centripetal) and gravity (tangential). Hence, the resultant force isn't purely towards the center, which is why the problem states that the resultant force does not always point towards the center.

Similarly, at the rightmost point, the velocity is upward (for clockwise motion), and gravity is acting downward. At the exact rightmost point, the velocity is purely upward, tangent to the circle. Gravity is acting downward, which is opposite to the direction of motion. Therefore, gravity here is providing a tangential deceleration, again contributing to a tangential component of acceleration. Tension here is horizontal towards the center (leftward at the rightmost point). Thus, again, the resultant force is a combination of the tension (centripetal) and gravity (tangential), resulting in a net force that is not purely centripetal.

Therefore, at the leftmost and rightmost positions, the resultant force on the weight is the vector sum of the tension force (directed towards the center of the circle) and the gravitational force (directed downward). Since these forces are perpendicular to each other, the resultant force vector will point in a direction that is a combination of centripetal (horizontal) and tangential (vertical) components. This resultant force is not purely towards the center because of the gravitational force acting downward, which introduces a tangential component. Hence, the net force is not centripetal at these points but has both centripetal and tangential components, leading to a resultant force that does not point directly towards the center of the circle.

But wait, let me make sure. Let's take the leftmost point: the tension is to the right (centripetal direction), gravity is down. The net force is the vector sum of these two. So the resultant force would be diagonally right and down. Similarly, at the rightmost point, tension is to the left, gravity is down, so net force is diagonally left and down. So neither of these resultant forces points directly towards the center of the circle. Therefore, the reason the resultant force doesn't always point towards the center is that, in vertical circular motion, gravity acts downward and is not aligned with the string's tension, introducing a tangential component to the net force at positions where the string is horizontal (leftmost and rightmost points). In contrast, at the top and bottom, gravity is either aligned with or opposite to the tension, leading to net forces that are centripetal (if we consider the bottom, where tension upwards minus weight provides centripetal; at the top, tension plus weight if any provides centripetal). Wait, but at the top, if the tension is insufficient, the weight might not complete the circle. But assuming it's in motion, then the net force at top and bottom is still towards the center, but at the sides, the net force is a combination of centripetal and tangential components.

Wait, to summarize: In vertical circular motion, due to gravity, the net force (resultant force) has both centripetal and tangential components except at the top and bottom, where gravity is along the vertical axis. At the leftmost and rightmost positions, the tension is horizontal, gravity is vertical, so their vector sum is not purely horizontal (towards the center), hence the resultant force points towards the center and downward, making it not purely centripetal but a combination.

Therefore, the answer would be that at the leftmost and rightmost positions, the resultant force is a combination of the tension force towards the center and the gravitational force downward. The resultant force does not always point towards the center because gravity introduces a tangential component to the net force at these positions.

I need to make sure there's no error here. Another way to think about it is, in uniform circular motion, the net force must be centripetal, but when the motion is non-uniform (speed changing), there is a tangential component. Since gravity causes the speed to change as the weight moves around the circle, the motion is non-uniform, hence the net force has tangential components. At leftmost and rightmost positions, the tangential component is vertical (due to gravity), so indeed the net force has both centripetal and tangential parts, resulting in a non-centripetal resultant direction.

Yes, that makes sense. Therefore, the conclusion is correct.
</think>

At the leftmost and rightmost positions of the weight's vertical circular motion, the **resultant force** is the vector sum of two forces:
1. **Tension** in the string, directed horizontally toward the center of the circle.
2. **Gravity** (weight), acting vertically downward.

Since these forces are perpendicular to each other, the resultant force vector points diagonally **toward the center and downward** at the leftmost position and diagonally **toward the center and downward** at the rightmost position (due to gravity's consistent downward pull). This resultant force has both a **centripetal component** (from tension) and a **tangential component** (from gravity).

Why the resultant force does not always point toward the center:
In vertical circular motion, gravity introduces a **tangential acceleration** component at positions where the string is horizontal (leftmost and rightmost points). This tangential component causes the speed of the weight to change (increase on the descending side and decrease on the ascending side), making the motion non-uniform. Consequently, the net force must account for both the centripetal force required for circular motion and the tangential force due to gravity. The resultant force vector therefore does not point purely toward the center but instead combines centripetal and tangential directions.

At the top and bottom of the circle, gravity aligns with the radial direction, resulting in a purely centripetal net force. However, at the sides (leftmost and rightmost points), gravity acts tangentially, leading to the non-centripetal resultant force."

Reasoning Strategies: ['self-verification', 'backtracking', 'synthesis', 'discussion', 'exploration', 'analysis', 'calculation', 'explanation', 'generalization']

Figure 12: Reasoning trace with high count of reasoning strategies, and annotated reasoning strategies.