NATURALREASONING: Reasoning in the Wild with 2.8M Challenging Questions

Weizhe Yuan[1,2] **Jane Yu**[1*] **Song Jiang**[1*] **Karthik Padthe**[1] **Yang Li**[1]
Ilia Kulikov[1] **Kyunghyun Cho**[2] **Dong Wang**[1] **Yuandong Tian**[1]
Jason Weston[1,2] **Xian Li**[1†]

[1]Meta [2]New York University

Abstract

Scaling reasoning capabilities beyond traditional domains such as math and coding is hindered by the lack of diverse and high-quality questions. To overcome this limitation, we introduce a scalable approach for generating diverse and challenging reasoning questions, accompanied by reference answers. We present NATURALREASONING, a comprehensive dataset comprising 2.8 million questions that span multiple domains, including STEM fields (e.g., Physics, Computer Science), Economics, Social Sciences, and more. We demonstrate the utility of the questions in NATURALREASONING through knowledge distillation experiments which show that NATURALREASONING can effectively elicit and transfer reasoning capabilities from a strong teacher model. Furthermore, we demonstrate that NATURALREASONING is also effective for unsupervised self-training using external reward models or self-rewarding. To foster future work, we publicly release NATURALREASONING at `https://huggingface.co/datasets/facebook/natural_reasoning`.

1 Introduction

Large language models (LLMs) have demonstrated increased reasoning capabilities [OpenAI et al., 2024, Guo et al., 2025]. These models are designed to devote more time to deliberation before responding, enabling them to tackle intricate tasks and solve more complex problems in science, coding, and mathematics. Such reasoning models are trained via large-scale reinforcement learning on tasks where the reward can be derived using rule-based verification. Existing reasoning datasets are often limited to narrow domains where the solutions are short and easy to verify, while the majority of reasoning problems across broader domains are open-ended reasoning. To bridge this gap, we introduce NATURALREASONING, a comprehensive dataset curated from pretraining corpora, comprising 2.8 million reasoning questions spanning various topics, including Mathematics, Physics, Computer Science, Economics & Business, etc. NATURALREASONING is compared to a wide range of reasoning datasets, showcasing its advantageous properties, in particular its diversity and difficulty.

NATURALREASONING possesses several desirable attributes as a dataset, serving multiple research purposes. Firstly, the questions are backtranslated from the pretraining corpora, ensuring that it represents **diverse** reasoning problems in the real world, as opposed to synthetic datasets derived from benchmark datasets like MetaMathQA [Yu et al.] and OpenMathInstruct-2 [Toshniwal et al.]. Secondly, it consists of both questions with **easy-to-verify** answers and those with **open-ended** solutions (e.g., theorem proving), providing a rich resource for developing learning algorithms to enhance LLMs' reasoning abilities across broader domains.Thirdly, we show that NATURALREASONING

* Equal contribution.
† Correspondence to: Xian Li <xianl@meta.com>.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

 https://doi.org/10.52202/085713-2246

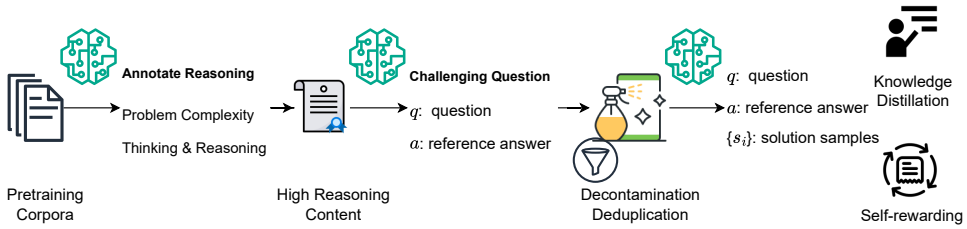


Figure 1: An overview of the data creation approach of NATURALREASONING.

poses more **difficult** reasoning problems than existing datasets. Its questions therefore provide an effective testbed for improving LLM reasoning—whether through knowledge distillation from a stronger teacher model or reinforcement learning with external and self-generated reward signals [Yuan et al.]. Lastly, the NATURALREASONING dataset complements existing reasoning datasets in terms of both **quality** and **quantity**.

Our contributions are threefold:

- We create a large-scale and high-quality reasoning dataset by using pretraining data and LLMs alone without extra human annotation. The dataset contains challenging questions which require deliberate thinking accompanied with reference answers. We release the NATURALREASONING dataset to the research community at `https://huggingface.co/datasets/facebook/natural_reasoning`.
- We show that NATURALREASONING is a highly performant dataset to enhance reasoning capabilities in post-training. Specifically, using questions from NATURALREASONING in distillation is more sample efficient than existing datasets.
- We investigate how NATURALREASONING supports self-training methods. Our results show that the questions in our dataset effectively enable self-learning, where self-rewarding techniques can yield performance comparable to some strong external reward models.

2 Data Collection

Backtranslating questions based on pretraining corpora has been shown to be a scalable approach to create instruction-following datasets [Li et al., 2023, Yue et al., 2024]. We take a similar approach to extract diverse and realistic reasoning questions grounded in pretraining data, i.e. we generate *grounded synthetic data*. An overview of the data creation approach is illustrated in Figure 1. A key differentiator of our approach is its emphasis on simplicity; we use LLMs throughout the entire cycle of synthetic data generation and curation, without any human annotation nor manual steps such as preprocessing to detect relevant documents and extracting questions with rule-based filtering.

2.1 Annotate Reasoning

Inspired by the meta-cognitive capabilities of state-of-the-art LLMs [Didolkar et al., 2024], we use an LLM to perform detailed annotation of pretraining documents to detect documents which contain sophisticated reasoning traces. We use the public pretraining corpora DCLM-baseline [Li et al., 2024b] and FineMath [Allal et al., 2025] as sources, which have demonstrated steeper scaling laws than alternative corpora. More specifically, given a document d from the pretraining corpus, we prompt an LLM to rate the content in d along multiple axes: Problem Completeness, Problem Complexity and Technical Depth, Technical Correctness and Accuracy, Thinking and Reasoning. The detailed prompt is provided in Appendix I. Empirically, we found that the model could analyze the document well and is able to follow the format requirement in the instruction.

2.2 Synthesize Questions and Answers

For documents which are identified with a high degree of reasoning (e.g., scored full points on axes of Problem Complexity and Technical Depth, Thinking and Reasoning), we further prompt an LLM to

Table 1: Comparison of four large publicly available reasoning datasets with NATURALREASONING. "Q" denotes "question", and question length is measured by the number of words.

Dataset	Domain	Q. #	Q. Len.	Question source	Response model
MetaMathQA	Math	395K	41±24	Rephrased by GPT-3.5-Turbo from GSM8K+MATH	GPT-3.5-Turbo
NuminaMath	Math	860K	48±32	Grade-level and competition-level math datasets	GPT-4o
OpenMath2	Math	607K	46±44	Synthesized by Llama3-405B-Inst from GSM8K+MATH	Llama3-405B-Inst
WebInstruct	Multi	13M	34±28	Recalled and Extracted from the Web	Mixtral-22B×8, Qwen-72B
NaturalReasoning	Multi	2.8M	55±21	Synthesized by Llama-3.3-70B-Inst grounded in the Web	Llama-3.3-70B-Inst

compose a self-contained and challenging reasoning question q based on the content in d. Different from existing work, which extracts questions appearing in the document [Yue et al., 2024], our approach allows us to synthesize more novel questions not directly contained in pretraining corpora. Then, we prompt the LLM to verify whether a correct reference answer a to the synthesized question q can be derived from d and, if possible, include it as a reference answer. Finally, for every question we generate an additional response with a relatively strong open-source model (Llama-3-70B-Instruct), which we later use as a teacher signal for knowledge distillation (see Section 5).

2.3 Question Deduplication and Decontamination

Deduplication Our deduplication process focuses on identifying and removing near-duplicate questions using locality-sensitive min-hashing at the word level. We apply a similarity threshold of 0.55 to filter out closely related variations, ensuring that questions with the same core reasoning task but different prompts are not redundantly included.

Decontamination We filter out questions that are similar to popular reasoning benchmarks including MATH [Hendrycks et al., 2021b], GPQA, MMLU-Pro [Wang et al., 2024] and MMLU-Stem [Hendrycks et al., 2021a]. The standard 13-gram decontamination method from EleutherAI's llm-eval-harness [Gao et al., 2024] is used to identify and remove 0.026% items from the dataset.

3 Data Analysis

We compare NATURALREASONING to the following representative existing datasets which were curated to boost reasoning capabilities.

MetaMathQA is created by bootstrapping mathematical questions from GSM8K and MATH, rewriting them from multiple perspectives to form a new dataset [Yu et al., 2024]. The responses are generated using GPT-3.5-Turbo.

NuminaMath is a comprehensive collection of 860K pairs of math problems and solutions [Li et al., 2024c]. The questions cover multiple sources including grade-level questions and competition problems. The solutions in NuminaMath dataset are generated or rewritten by GPT-4o.

OpenMathInstruct-2 is a collection of 14M synthesized math questions and solutions, based on GSM8K and MATH [Toshniwal et al., 2024]. The solutions are generated by Llama3.1-405B-Instruct [Grattafiori et al., 2024] and curated through majority vote on the final answer.

WebInstruct recalls relevant documents from Common Crawl using a fastText model trained on a diverse seed dataset of quiz websites. It then extracts question-answer pairs contained in recalled web pages and uses LLMs (Qwen-72B [Bai et al., 2023], Mixtral-8×7B [Jiang et al., 2024]) to refine the answer [Yue et al., 2024].

In addition, we compare models trained on NATURALREASONING with those trained on the OpenThoughts dataset [Guha et al., 2025], a recent open-source collection designed for reasoning in math, code, and science. As shown in Appendix G, models trained on NATURALREASONING

https://doi.org/10.52202/085713-2246

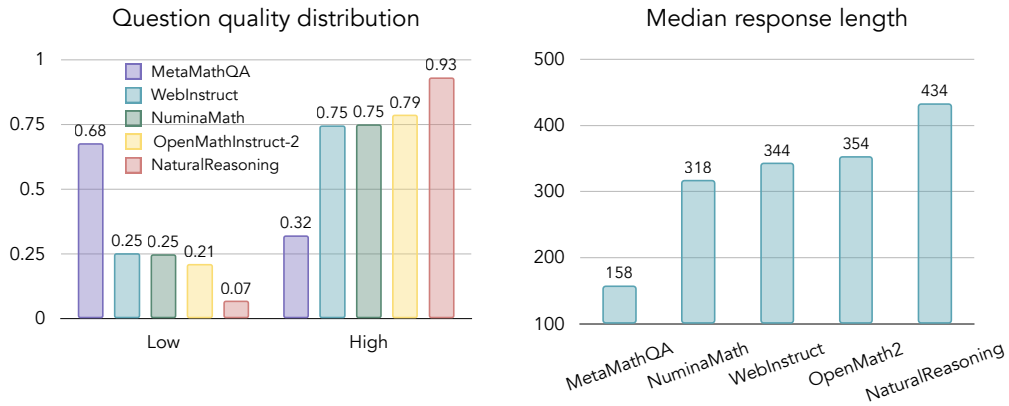


Figure 2: **Left**: Question quality distribution based on LLM annotations: Low (0-6), High (7-10). **Right**: Median response length (in words) of Llama3.3-70B-Instruct responses across all datasets.

achieve better performance across general reasoning benchmarks, demonstrating its broader coverage and stronger generalization compared to OpenThoughts.

3.1 Basic Statistics

We present a comparison of key dataset statistics in Table 1. Most large open reasoning datasets primarily focus on the math domain, with datasets such as OpenMathInstruct-2, NuminaMath, and MetaMathQA containing only math-related questions. In contrast, NATURALREASONING covers reasoning problems from more diverse domains. Additionally, NATURALREASONING consists of 2.8M **unique** questions, significantly larger than OpenMathInstruct-2 (607K), NuminaMath (860K), and MetaMathQA (395K), though smaller than WebInstruct (13M).

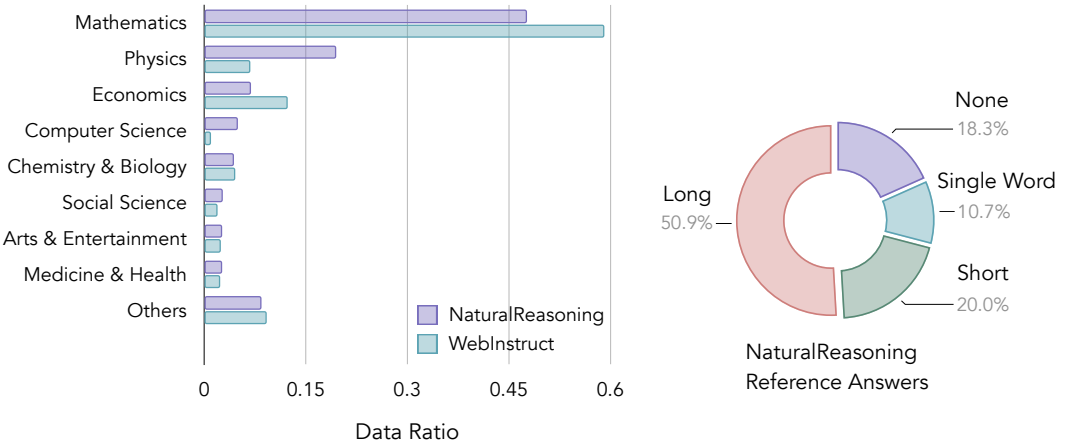
With an average length of 55 words, questions in NATURALREASONING are notably longer than those in OpenMathInstruct-2 (46), WebInstruct (34), NuminaMath (48), and MetaMathQA (41). Longer questions embed richer context and multi-step requirements, demanding deeper reasoning. Coupled with our varied, web-grounded sources, this added complexity sets NATURALREASONING apart as a uniquely challenging dataset.

3.2 Question Quality and Difficulty

Quality We evaluate question quality in NATURALREASONING with both automatic and human judgments. Three strong LLMs (DeepSeek-R1-Distill-Qwen-32B, Qwen2.5-72B-Instruct, Llama-3-70B-Instruct) independently score every question on a 0–10 scale reflecting solvability and completeness. For each comparison dataset, we randomly sample 10% of its questions and apply the same prompt. Scores of 0–6 are labeled low quality and 7–10 high quality; a question is deemed high quality when at least two models assign a high score. As shown in Figure 2, NATURALREASONING contains the highest fraction of high-quality questions (93%), surpassing the next-best dataset (79%). To corroborate these results, we conduct a small human study: two expert annotators independently score 100 randomly selected questions from each dataset, and we average their ratings. The pattern holds—NATURALREASONING achieves the top mean score (6.45) versus 5.92 for the second best. Full details are provided in Appendix D.

Difficulty To estimate question difficulty, we leverage a strong LLM to generate responses and use response length as a proxy, as longer chain-of-thoughts typically correspond to more complex questions. Specifically, we randomly selected 10% of questions from each dataset, and employ Llama3.3-70B-Instruct to generate responses for each question. As is shown in Figure 2, NATURALREASONING exhibits the longest median response length (434 words), significantly surpassing all other datasets. This suggests that our dataset contains more intricate and reasoning-demanding questions compared to existing open reasoning datasets.

https://doi.org/10.52202/085713-2246



(a) Question topic distributions.

(b) Reference answer lengths in NATURALREASONING.

Figure 3: **Left**: Topic distributions of NATURALREASONING and WebInstruct. NATURALREASONING generally shows equivalent or even greater coverage on non-Math topics like Computer Science and Physics. **Right**: Distribution of reference answer lengths in NATURALREASONING, showing that the majority of questions have long reference answers ($\geq$10 words).

3.3 Question Diversity

In addition to being difficult, questions in NATURALREASONING are also diverse. We analyze diversity of the questions in terms of question similarity and the topics, and compare to WebInstruct, an existing dataset covering multiple domains.

Embedding Clustering We use an off-the-shell sentence encoder[3] to generate embeddings for questions in WebInstruct and NATURALREASONING. We then apply UMAP [McInnes et al., 2018] to project the high-dimensional embeddings into a 2D space, followed by K-means clustering [Wu, 2012] to identify distinct groups. We use Mixtral-8B to assign high-level labels to these clusters, which is prompted with a few examples from each cluster. As Figure 5 shows, NATURALREASONING contains a more diverse and dense representation of non-mathematical topics, including Physics, Chemistry, Computer Science, and Law, besides Math. In contrast, WebInstruct is primarily skewed toward mathematical content, highlighting the broader topic coverage of NATURALREASONING.

Classifier Categorization To estimate the topic distribution, a multi-class topic classifier is used to classify each question into 16 knowledge classes. The class labels are motivated by Wikipedia academic disciplines[4]. Figure 3a shows that NATURALREASONING is complementary to WebInstruct, where NATURALREASONING has better coverage on non-Math topics especially Physics, Computer Science, Social Science, etc.

3.4 Reference Answer Analysis

Among the 2.8 million questions we synthesized, 81.68% have reference answers which could be derived from pretraining data. The distribution of reference answer lengths is illustrated in Figure 3b, with single-word answers accounting for 10.7%, short answers (2–9 words) making up 20.0%, and long answers ($\geq$10 words) constituting the majority at 50.9%.

We provide examples of questions with single-word, short, and long answers in Appendix B. In general, we found that questions with single word answers typically involve numerical, factual, or definitional queries, while questions with long answers demand more free-form in-depth analysis. For questions with a long answer, the extracted reference answer is typically a short summary content

[3] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[4] https://en.wikipedia.org/wiki/Outline_of_academic_disciplines

https://doi.org/10.52202/085713-2246

from the original documents or useful clues to answer the question. While reference answers may contain some noise, we demonstrate their utility in Appendix F for both filtering training data in knowledge distillation and enabling reinforcement learning with verifiable rewards (RLVR) [Lambert et al., 2025].

4 Experimental Setup

We highlight the efficacy of NATURALREASONING in two settings: (1) **Knowledge distillation**, and (2) **Unsupervised Self-Training**. For (1), we evaluate whether NATURALREASONING enables steeper scaling than existing datasets when distilling reasoning capabilities to a student model via supervised finetuning (Section 5). We experiment with different model families such as Llama3.1-8B and Qwen2.5-7B. Specifically we show that questions from NATURALREASONING are very effective at distilling long chain-of-thoughts from reasoning models and compare it to manually curated questions such as LIMO [Ye et al., 2025] and S1K [Muennighoff et al., 2025] (Section 6). To demonstrate (2), we evaluate how well NATURALREASONING supports self-training either through a strong external reward model or self-rewarding mechanisms [Yuan et al., 2024] (Section 7).

Evaluation We evaluate our models on a diverse set of benchmarks that encompass both math and science reasoning: MATH, GPQA, GPQA-Diamond [Rein et al., 2024], and MMLU-Pro. In Appendix H, We also show NATURALREASONING's utility for broader NLP tasks (e.g., writing). To ensure a fair and consistent comparison, we adopt a zero-shot evaluation setting across all trained models. For inference we use vllm [Kwon et al., 2023] and employ greedy decoding to maintain determinism and eliminate variability introduced by stochastic generation. Unless mentioned otherwise, we report accuracy averaged over the last three saved model checkpoints during training.

5 Steeper Scaling with Challenging and Diverse questions

Our hypothesis is that challenging and diverse questions which require thinking and reasoning are more sample efficient for post-training. To verify this, we run supervised finetuning (SFT) starting from a base model, and evaluate overall performance across MATH, GPQA, and MMLU-Pro.

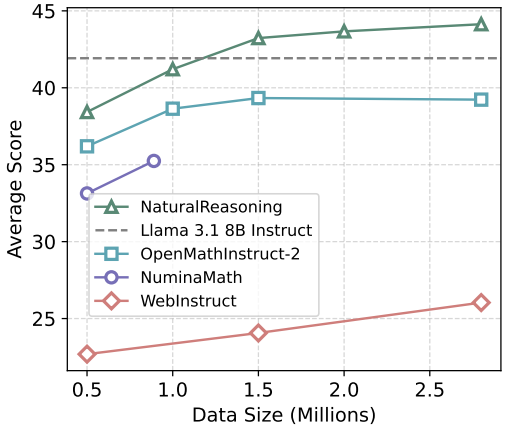
We fine-tuned the Llama3.1-8B-Base model and Qwen2.5-7B using fairseq2 training recipes [Balioglu, 2023], exploring the impact of varying dataset sizes. Specifically, we trained models on our dataset and the comparison datasets introduced in Table 1, and evaluated their average performance across three benchmarks: MATH, GPQA, and MMLU-Pro. For all datasets, we train 3 epochs for each data size, using batch size 128, learning rate $5e^{-6}$, with cosine learning rate schedule where final learning rate is 1% of peak learning rate.

5.1 Results

The scaling trends for Llama3.1-8B-Base model plotted by averaging performance on the three benchmarks are shown in Figure 4a, and Figure 4b provides a detailed breakdown of model performance across different dataset sizes and benchmarks. The scaling trends for Qwen2.5-7B model also show the superiority of NATURALREASONING and we put the results in Appendix E.

NATURALREASONING is significantly more sample-efficient than existing reasoning datasets. As shown in Figure 4a, models trained on NATURALREASONING require fewer training examples to achieve superior performance. With just 1.5 million training examples, the model trained on NATURALREASONING already outperforms Llama3.1-8B-Instruct, which was extensively tuned for instruction-following with more data [Grattafiori et al., 2024]. In contrast, other datasets, including OpenMathInstruct-2 and WebInstruct, fail to surpass Llama3.1-8B-Instruct even when trained on 2.8 million data. Each NATURALREASONING sample therefore provides denser, more effective reasoning supervision, making it the most data-efficient choice for boosting model reasoning performance.

Math-specific datasets like OpenMathInstruct-2 excel at math reasoning but fail to generalize beyond math. A closer look at Figure 4b reveals that OpenMathInstruct-2 consistently achieves the highest scores on the MATH benchmark, with performance increasing from 50.83 (500K) to 59.25 (2.8M). This confirms that OpenMathInstruct-2 is well-optimized for pure math reasoning. However,



(a) Average performance across MATH, GPQA, and MMLU-Pro when using varying sizes of data for training Llama3.1-8B-Base. SFT with 1.5 million examples from NATURALREASONING is able to surpass Llama3.1-8B-Instruct.

(b) Performance breakdown by benchmark, where the highest accuracy per data size is bolded.

SFT Dataset	500K	1M	1.5M	2.8M
		MATH		
WebInstruct	9.60	–	11.65	14.46
NuminaMath	49.53	–	–	–
OpenMathInstruct-2	**50.83**	**54.58**	**56.47**	**59.25**
NaturalReasoning	45.17	48.55	52.49	55.55
		GPQA		
WebInstruct	**29.02**	–	25.37	26.12
NuminaMath	13.84	–	–	–
OpenMathInstruct-2	25.60	27.31	27.23	26.45
NaturalReasoning	26.64	**29.91**	**31.77**	**30.13**
		MMLU-Pro		
WebInstruct	29.44	–	35.17	37.54
NuminaMath	42.36	–	–	–
OpenMathInstruct-2	32.16	34.03	34.30	31.99
NaturalReasoning	**43.47**	**45.16**	**45.43**	**46.71**

Figure 4: Scaling results for Llama3.1-8B-Base model.

its performance on GPQA and MMLU-Pro is significantly weaker, where GPQA accuracy plateaus around 27–26 as dataset size increases, and MMLU-Pro accuracy fluctuates without significant improvement. This suggests that while OpenMathInstruct-2 provides strong supervision in math reasoning, it lacks the diversity required to generalize to broader scientific reasoning tasks.

Some datasets show diminishing returns as training data increases, highlighting potential inefficiencies in data composition. While scaling up dataset size generally improves performance, datasets like WebInstruct and OpenMathInstruct-2 exhibit inconsistent or plateauing performance trends. For example, WebInstruct's GPQA performance peaks at 500K (29.02) but drops at 1.5M (25.37) and only marginally improves at 2.8M (26.12). Similarly, OpenMathInstruct-2's GPQA accuracy fluctuates with increased training data, suggesting that simply adding more data does not always lead to better reasoning abilities. These observations imply that data quality and diversity matter more than data volume when training models for complex reasoning.

6 Eliciting Long Chain-of-Thought

In addition to the emergence of OpenAI-o1 and DeepSeek-R1, several studies suggest that simpler tasks require fewer steps while complex tasks benefit significantly from longer CoTs [Jin et al., 2024]. When encouraging the model to think for longer, they are able to solve questions that they previously could not. Motivated by this, we investigate whether the questions in NATURALREASONING are complex enough to benefit from longer CoTs from a stronger reasoning model. We do so by distilling Deepseek-R1 responses to Llama3.3-70B-Instruct.

We randomly sampled 1K questions from NATURALREASONING and used SGLang [Zheng et al., 2023] to prompt DeepSeek-R1, generating one response per question. Resulting response lengths range from 745 to 14.6K tokens with an average length of 4430 tokens. We then supervised finetune Llama-3.3-70B-Instruct on this set and compare its performance against training on two strong, heavily curated datasets: s1K-1.1 [Muennighoff et al., 2025] and LIMO [Ye et al., 2025]. Both datasets underwent multiple filtering stages to ensure that their questions are high-quality, diverse, and challenging; for consistency, all responses were generated with DeepSeek-R1. To keep the evaluation consistent with the setting used in Guo et al. [2025], we report pass@1 averaged across $n = 24$ samples. Each sample is generated using temperature=0.6, top_p=0.95.

The results in Table 2 show that a randomly selected subset of 1k questions from NATURALREASONING matches—even slightly exceeds—the performance obtained on datasets that underwent several

https://doi.org/10.52202/085713-2246

Table 2: Pass@1 of Llama-3.3-70B-Instruct after distilling DeepSeek-R1 responses. We compare the performance of random selection from NATURALREASONING with curated datasets such as s1K-1.1[Muennighoff et al., 2025] and LIMO[Ye et al., 2025], as well as the scaling effect of NATURALREASONING.

	Training size	GPQA-Diamond	MMLU-Pro	MATH-500	Average
Llama3.3-70B-Instruct	0	50.5	70.5	77.0	66.0
LIMO	817	56.5	76.8	86.6	73.3
s1K-1.1	1,000	62.7	77.4	86.6	75.6
NATURALREASONING	1,000	63.5	78.0	86.2	75.9
NATURALREASONING	10,000	65.6	78.4	87.4	77.1
NATURALREASONING	100,000	**67.3**	**79.5**	89.8	78.9
DeepSeek-R1-Distill-Llama-70B	800,000	65.2	78.5	**94.5**	79.4

rounds of meticulous filtering and curation. This parity underscores that NATURALREASONING contains questions that are diverse, challenging, and of consistently high quality.

To examine the impact of scale, we expanded the random sample size from 1K to 10K and finally to 100K NATURALREASONING questions; the corresponding results are also presented in Table 2. Performance increases monotonically across every benchmark, providing further evidence that enlarging the slice of NATURALREASONING delivers substantial gains precisely because the added questions maintain the same high standard of quality. Moreover, fine-tuning Llama-3.3-70B-Instruct on just 100K randomly sampled questions from NATURALREASONING brings the model close to DeepSeek-R1-Distill-Llama-70B, which was trained on 800K examples. Despite using only one-eighth the data, our model outperforms DeepSeek-R1-Distill-Llama-70B on GPQA-Diamond and MMLU-Pro, and falls only slightly behind on MATH-500. This further attests to both the scalability and the intrinsic quality of NATURALREASONING.

7 Unsupervised Self-Training

Since open-ended reasoning questions are difficult to evaluate using exact match with reference answers, we explore whether our dataset can facilitate self-training through either strong external reward models or self-reward mechanisms [Yuan et al.]

To test the effectiveness of self-training without confounding factors from distribution shift, we evaluate on GPQA-Diamond as test set, and use the remaining questions from GPQA as seeds to retrieve similar questions from NATURALREASONING. We curated 15,000 questions in total, which we refer to as `SelfTrain-15k`. More details are in Section C.2. Models trained on this subset can still be evaluated on the GPQA-Diamond test set as those questions are not used for data selection.

We verify unsupervised self-training under two different training method: Rejection-based sampling Fine-Tuning (RFT) and Direct Preference Optimization (DPO) [Rafailov et al., 2023], focusing on the effectiveness of different reward scoring strategies. Each approach relies on sampling 32 candidate responses per question, followed by selecting responses based on reward scores. RFT employs rejection sampling, selecting the highest-scoring response for SFT training, while DPO constructs training pairs using both the highest and lowest-scoring responses. For external reward models, we consider `Qwen2.5-Math-RM-72B` [Yang et al., 2024] and `INF-ORM-Llama3.1-70B`[5]. In addition, we explore a self-rewarding framework where the model evaluates and assigns rewards to its own generated responses. Specifically, we consider the following self-rewarding strategies:

Self-consistency: Inspired by prior work such as Prasad et al. [2024], the best response is selected based on response frequency, while the worst response is chosen randomly. To determine frequencies, we extract final answers formatted as `\boxed{X}` and compute their occurrence counts. Responses without a clearly extractable final answer are filtered out.

Self-scoring: The model receives the question and candidate response in a single prompt and is asked to assess whether the response is valid. We define the reward as the log-probability difference between the judgements "yes" and "no". The full prompt is in Figure 11.

[5]https://huggingface.co/infly/INF-ORM-Llama3.1-70B

Table 3: Unsupervised self-training results. We employ RFT and DPO training of Llama3.1-8B-Instruct, using various reward scoring strategies.

Model	GPQA-Diamond	MMLU-Pro	Average
Llama3.1-8B-Instruct	31.82	49.79	40.81
RFT training using external reward model & self-reward			
INF-ORM-Llama3.1-70B	32.66	50.95	41.81
Qwen2.5-Math-RM-72B	34.18	49.84	42.01
Self-consistency	34.18	49.83	41.91
Self-score	34.34	50.36	42.35
Self-score-filtered	**35.02**	50.06	42.54
DPO training using external reward model & self-reward			
INF-ORM-Llama3.1-70B	33.50	**52.74**	43.12
Qwen2.5-Math-RM-72B	30.13	49.17	39.65
Self-consistency	30.81	48.60	39.71
Self-score	34.34	52.11	43.22
Self-score-filtered	**35.02**	52.31	**43.67**

Self-scoring with filtering: on top of self-scoring, when applying RFT or DPO, we introduce an additional filtering mechanism. Specifically, for RFT, if the highest-ranked response has a self-score below zero, it is discarded. For DPO, if the preferred response in a pair has a self-score below zero, the pair is removed from training.

We train Llama3.1-8B-Instruct using RFT data and DPO data constructed through these methods. We use learning rate of $1e^{-6}$, batch size of 64, and train for three epochs, with checkpoints saved every 50 steps. We report test performance on GPQA-Diamond and MMLU-Pro in Table 3.

7.1 Results

Self-training improves performance over the baseline. Llama3.1-8B-Instruct, serving as the baseline, achieves an average score of 40.81 across GPQA-Diamond and MMLU-Pro. Almost all self-training methods lead to improvements, demonstrating the effectiveness of fine-tuning on high-quality model-generated responses.

Self-reward methods are highly competitive, often surpassing external reward models. While using external reward models, such as INF-ORM-Llama3.1-70B, could outperform the baseline, self-reward methods achieve comparable or even superior results. Notably, self-score-filtered SFT and self-score-filtered DPO deliver the best performance on GPQA-Diamond (35.02), with self-score-filtered DPO achieving the highest overall score (43.67). These results highlight that self-reward mechanisms can effectively guide self-training without relying on external reward models.

Self-score filtering further enhances performance by improving training data quality. Among self-reward methods, applying simple filtering improves results across both RFT and DPO. In RFT, self-score-filtered (42.54 AVG) outperforms unfiltered self-scoring (42.35 AVG), while in DPO, self-score-filtered (43.67 AVG) surpasses unfiltered self-scoring (43.22 AVG). This suggests that filtering out low-confidence responses strengthens self-training by reducing noise in the training data.

8 Related Work

Synthetic Reasoning Data. Synthetic data has emerged as a promising solution for improving performance. Some approaches bootstrap new data from existing data (e.g., STaR [Zelikman et al., 2022] augments with new CoT rationales and MetaMath [Yu et al.] rewrites the questions in MATH and GSM8K in several ways), but these techniques rely on the existence of a high-quality dataset. Other techniques such as that of OpenMathInstruct-2 [Toshniwal et al.], Xwin-Math [Li et al., 2024a], and Self-Instruct [Wang et al., 2023] generate new data from only a few seed examples using an LLM but scaling to new domains remains a significant challenge. MMIQC [Liu et al.] parses QA pairs

 https://doi.org/10.52202/085713-2246

from Mathematics Stack Exchange, using the highest-ranked answer, but few measures are taken to curate for quality and the resulting dataset is also specific to the math domain. Similar to our work, WebInstruct [Yue et al., 2024] harvests question-answer pairs from pre-training corpora and spans multiple domains, but is dependent on carefully crafted rule-based filters.

Unsupervised Self-training Most prior works typically depend on human-annotated (gold) final answers [Zelikman et al., 2022, Chen et al., 2024, Pang et al., 2024] or the use of an external reward model [Singh et al., 2024, Dong et al., 2023]. However, manually annotating or verifying final answers is particularly resource-intensive for complex, multi-step problems and training effective reward models for reasoning often requires human evaluation of LLM outputs [Cobbe et al., 2021, Uesato et al., 2022, Lightman et al., 2023], making it similarly costly. Like works such as She et al. [2024], Yuan et al. [2024], Rosset et al. [2024], Tran et al. [2023], our work explores self-training in the absence of gold labels and does not limit itself to questions with short, easily verifiable targets.

9 Conclusion

We present NATURALREASONING, a dataset of 2.8 million questions for enhancing LLM reasoning capabilities. Our questions are challenging, requiring more deliberate thinking than existing datasets. The dataset covers diverse reasoning problems across multiple domains including math, physics, computer science, economics, social sciences, etc. Using questions from NATURALREASONING in distillation experiments, we observe consistent improvement on reasoning benchmarks when scaling the data size. We also demonstrate that NATURALREASONING is effective for enabling LLM unsupervised self-training using external reward models or self-rewarding.

Limitation & Impact Statement

Although our study already validates NATURALREASONING 's value for large-scale offline training—covering supervised distillation and preference-based self-training (RFT, DPO)—we also conduct preliminary experiments using online RL with verifiable rewards with General Verifier, which show promising gains even with limited training. A more systematic exploration of reinforcement learning paradigms, including alternative reward models and scaling strategies remains natural extensions for future work. This paper seeks to improve reasoning capabilities of large language models through leveraging pretraining corpora. While our efforts are focused on curating high-quality, diverse data, models trained using this data may exhibit undesirable behavior not examined in our work. Therefore, comprehensive evaluation would be needed to evaluate and address any potential pre-existing or existing biases in LLMs which leverage this data.

References

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. Smollm2: When smol goes big – data-centric training of a small language model, 2025. URL https://arxiv.org/abs/2502.02737.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report, 2023. URL https://arxiv.org/abs/2309.16609.

Can Balioglu. fairseq2, 2023. URL http://github.com/facebookresearch/fairseq2.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=O4cHTxW9BS.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Aniket Didolkar, Anirudh Goyal, Nan Rosemary Ke, Siyuan Guo, Michal Valko, Timothy Lillicrap, Danilo Rezende, Yoshua Bengio, Michael Mozer, and Sanjeev Arora. Metacognitive capabilities of llms: An exploration in mathematical problem solving. *arXiv preprint arXiv:2405.12205*, 2024.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, KaShun Shum, and Tong Zhang. RAFT: Reward ranked finetuning for generative foundation model alignment. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=m7p5O7zblY.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 07 2024. URL https://zenodo.org/records/12608602.

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,

 https://doi.org/10.52202/085713-2246

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. URL `https://arxiv.org/abs/2407.21783`.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer, Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, Wanjia Zhao, John Yang, Shreyas Pimpalgaonkar, Kartik Sharma, Charlie Cheng-Jie Ji, Yichuan

Deng, Sarah Pratt, Vivek Ramanujan, Jon Saad-Falcon, Jeffrey Li, Achal Dave, Alon Albalak, Kushal Arora, Blake Wulfe, Chinmay Hegde, Greg Durrett, Sewoong Oh, Mohit Bansal, Saadia Gabriel, Aditya Grover, Kai-Wei Chang, Vaishaal Shankar, Aaron Gokaslan, Mike A. Merrill, Tatsunori Hashimoto, Yejin Choi, Jenia Jitsev, Reinhard Heckel, Maheswaran Sathiamoorthy, Alexandros G. Dimakis, and Ludwig Schmidt. Openthoughts: Data recipes for reasoning models, 2025. URL https://arxiv.org/abs/2506.04178.

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, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021a.

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. *NeurIPS*, 2021b.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts, 2024. URL https://arxiv.org/abs/2401.04088.

Mingyu Jin, Qinkai Yu, Dong Shu, Haiyan Zhao, Wenyue Hua, Yanda Meng, Yongfeng Zhang, and Mengnan Du. The impact of reasoning step length on large language models. *arXiv preprint arXiv:2401.04925*, 2024.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626. ACM, October 2023. doi: 10.1145/3600006.3613165. URL http://dx.doi.org/10.1145/3600006.3613165.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training, 2025. URL https://arxiv.org/abs/2411.15124.

Chen Li, Weiqi Wang, Jingcheng Hu, Yixuan Wei, Nanning Zheng, Han Hu, Zheng Zhang, and Houwen Peng. Common 7b language models already possess strong math capabilities. *arXiv preprint arXiv:2403.04706*, 2024a.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, et al. Datacomp-lm: In search of the next generation of training sets for language models. *arXiv preprint arXiv:2406.11794*, 2024b.

Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. Numinamath. [https://huggingface.co/AI-MO/NuminaMath-CoT](https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf), 2024c.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Omer Levy, Luke Zettlemoyer, Jason Weston, and Mike Lewis. Self-alignment with instruction backtranslation. *arXiv preprint arXiv:2308.06259*, 2023.

 https://doi.org/10.52202/085713-2246

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Haoxiong Liu, Yifan Zhang, Yifan Luo, and Andrew C Yao. Augmenting math word problems via iterative question composing. In *ICLR 2024 Workshop on Navigating and Addressing Data Problems for Foundation Models*.

Xueguang Ma, Qian Liu, Dongfu Jiang, Ge Zhang, Zejun Ma, and Wenhu Chen. General-reasoner: Advancing llm reasoning across all domains. *arXiv:2505.14652*, 2025. URL `https://arxiv.org/abs/2505.14652`.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

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. URL `https://arxiv.org/abs/2501.19393`.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai,

Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. Openai o1 system card, 2024. URL https://arxiv.org/abs/2412.16720.

Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. Iterative reasoning preference optimization. *arXiv preprint arXiv:2404.19733*, 2024.

Archiki Prasad, Weizhe Yuan, Richard Yuanzhe Pang, Jing Xu, Maryam Fazel-Zarandi, Mohit Bansal, Sainbayar Sukhbaatar, Jason Weston, and Jane Yu. Self-consistency preference optimization. *arXiv preprint arXiv:2411.04109*, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.18290.

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. URL https://openreview.net/forum?id=Ti67584b98.

Corby Rosset, Ching-An Cheng, Arindam Mitra, Michael Santacroce, Ahmed Awadallah, and Tengyang Xie. Direct nash optimization: Teaching language models to self-improve with general preferences. *arXiv preprint arXiv:2404.03715*, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Shuaijie She, Shujian Huang, Wei Zou, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. MAPO: Advancing multilingual reasoning through multilingual alignment-as-preference optimization. *arXiv preprint arXiv:2401.06838*, 2024.

Avi Singh, John D Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J Liu, James Harrison, Jaehoon Lee, Kelvin Xu, Aaron T Parisi, Abhishek Kumar, Alexander A Alemi, Alex Rizkowsky, Azade Nova, Ben Adlam, Bernd Bohnet, Gamaleldin Fathy Elsayed, Hanie Sedghi, Igor Mordatch, Isabelle Simpson, Izzeddin Gur, Jasper Snoek, Jeffrey Pennington, Jiri Hron, Kathleen Kenealy, Kevin Swersky, Kshiteej Mahajan, Laura A Culp, Lechao Xiao, Maxwell Bileschi, Noah Constant, Roman Novak, Rosanne Liu, Tris Warkentin, Yamini Bansal, Ethan Dyer, Behnam Neyshabur, Jascha Sohl-Dickstein, and Noah Fiedel. Beyond human data: Scaling self-training for problem-solving with language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=lNAyUngGFK. Expert Certification.

M-A-P Team, Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, Kang Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, Chujie Zheng, Kaixin Deng, Shian Jia, Sichao Jiang, Yiyan Liao, Rui Li, Qinrui Li, Sirun Li, Yizhi Li, Yunwen Li, Dehua Ma, Yuansheng Ni, Haoran Que, Qiyao Wang, Zhoufutu Wen, Siwei Wu, Tianshun Xing, Ming Xu, Zhenzhu Yang, Zekun Moore Wang, Junting Zhou, Yuelin Bai, Xingyuan Bu, Chenglin Cai, Liang Chen, Yifan Chen, Chengtuo Cheng, Tianhao Cheng, Keyi Ding, Siming Huang, Yun Huang, Yaoru Li, Yizhe Li, Zhaoqun Li, Tianhao Liang, Chengdong Lin, Hongquan Lin, Yinghao Ma, Tianyang Pang, Zhongyuan Peng, Zifan Peng, Qige Qi, Shi Qiu, Xingwei Qu, Shanghaoran Quan, Yizhou Tan, Zili Wang, Chenqing Wang, Hao Wang, Yiya Wang, Yubo Wang, Jiajun Xu, Kexin Yang, Ruibin Yuan, Yuanhao Yue, Tianyang Zhan, Chun Zhang, Jinyang Zhang, Xiyue Zhang, Xingjian Zhang, Yue Zhang, Yongchi Zhao, Xiangyu Zheng, Chenghua Zhong, Yang Gao, Zhoujun Li, Dayiheng Liu, Qian Liu, Tianyu Liu, Shiwen Ni, Junran Peng, Yujia Qin, Wenbo Su, Guoyin Wang, Shi Wang, Jian Yang, Min Yang, Meng Cao, Xiang Yue, Zhaoxiang Zhang, Wangchunshu Zhou, Jiaheng Liu, Qunshu Lin, Wenhao Huang, and Ge Zhang. Supergpqa: Scaling llm evaluation across 285 graduate disciplines, 2025. URL https://arxiv.org/abs/2502.14739.

Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisacanin, Alexan Ayrapetyan, and Igor Gitman. Openmathinstruct-2: Accelerating ai for math with massive open-source instruction data. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*.

 https://doi.org/10.52202/085713-2246

Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisacanin, Alexan Ayrapetyan, and Igor Gitman. Openmathinstruct-2: Accelerating ai for math with massive open-source instruction data. *arXiv preprint arXiv:2410.01560*, 2024.

Hoang Tran, Chris Glaze, and Braden Hancock. Iterative DPO alignment. Technical report, Snorkel AI, 2023.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, 2023.

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.

Junjie Wu. *Advances in K-means clustering: a data mining thinking*. Springer Science & Business Media, 2012.

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

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

Longhui Yu, Weisen Jiang, Han Shi, YU Jincheng, Zhengying Liu, Yu Zhang, James Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. In *The Twelfth International Conference on Learning Representations*.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL `https://openreview.net/forum?id=N8N0hgNDRt`.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E Weston. Self-rewarding language models. In *Forty-first International Conference on Machine Learning*.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E Weston. Self-rewarding language models. In *Forty-first International Conference on Machine Learning*, 2024. URL `https://openreview.net/forum?id=0NphYCmgua`.

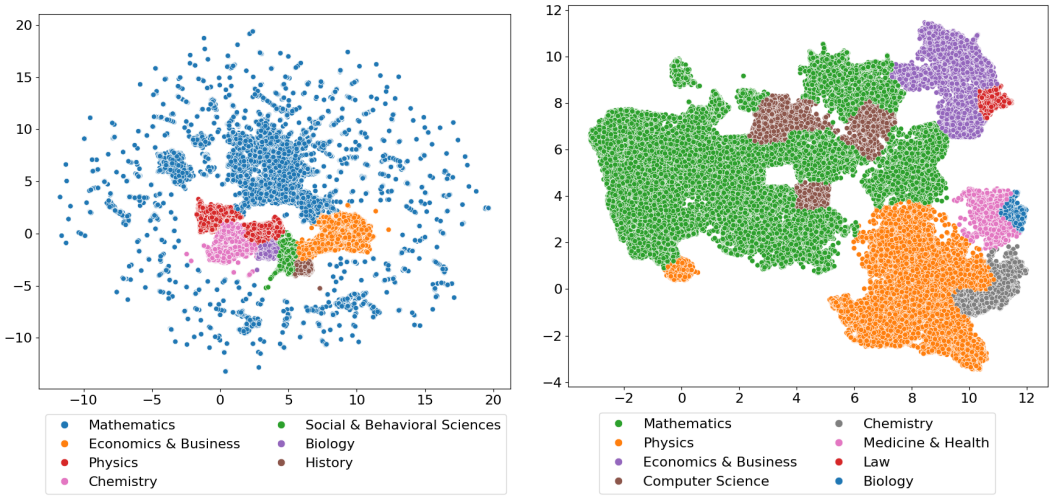
Xiang Yue, Tuney Zheng, Ge Zhang, and Wenhu Chen. Mammoth2: Scaling instructions from the web. *Advances in Neural Information Processing Systems*, 2024.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*, 2022.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. Sglang: Efficient execution of structured language model programs, 2023.

A Clustering Results

We present the results of our clustering in Figure 5. The procedure for producing this clustering is described in Section 3.3.



(a) WebInstruct (b) NATURALREASONING

Figure 5: Topic clustering of WebInstruct and NATURALREASONING.

B Example questions

Example questions with single word answer, short answer, long answer are shown in Table 4.

C Data Creation Details

C.1 Generation

We use vllm for all generations. For annotating documents and synthesizing questions, we use greedy decoding (i.e. temperature=0). For response generation for each question in NATURALREASONING, we use temperature=0.7 top_p=0.9. Responses used in unsupervised self-training experiments are sampled using temperature={0, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.2} to encourage response diversity.

C.2 SelfTrain-15k Curation

Specifically, for each question in the non-Diamond subset of GPQA, we retrieve the 1,000 most similar questions from NATURALREASONING, which were already decontaminated against the entire GPQA datset. Similarity is computed using cosine similarity between two question embeddings. After obtaining this candidate set, we apply deduplication and perform clustering, grouping the questions into 15,000 clusters. From each cluster, we select the questions closest to the cluster center, ensuring a diverse and representative dataset for downstream science reasoning tasks. This process resulted in a pool of 15,000 questions, which we refer to as `SelfTrain-15k`.

D Human Evaluation of Question Quality

To assess question quality reliably, we also conducted a human evaluation by randomly sampling 100 questions from each dataset. Two expert annotators independently rated each question, and we used the average of their scores as the final quality measure. The results, shown in Table 5, confirm that the NATURALREASONING dataset consistently produces higher-quality questions.

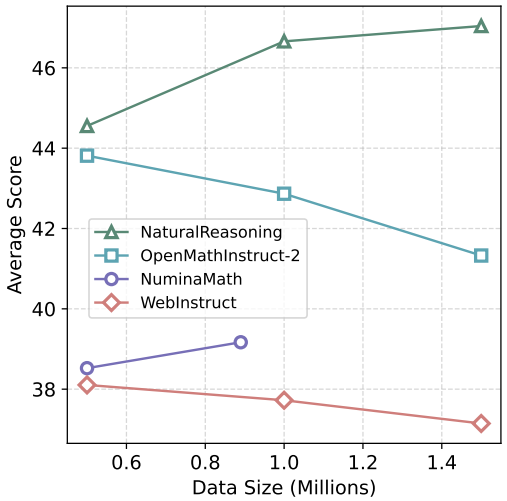
https://doi.org/10.52202/085713-2246

Table 4: Example questions with single word, short, and long reference answers.

Example questions with single word answer
(1) You have 5 cards with letters and numbers on either side. The rule is: If the letter on one side of a card is a vowel, the number on the other side is odd. What is the minimum number of cards you need to turn over to prove or disprove this rule? (2) What is the approximate number of grapes required to produce one bottle of wine, considering the conversion factors and variability in grape yields, and how does this relate to the overall wine production process? (3) A company is facing financial difficulties and is in danger of not meeting its obligations. In an effort to secure a bridge loan, the company's management decides to manipulate its financial statements by not reversing cancelled orders, thereby overstating its accounts receivable. This action allows the company to collateralize the loan and secure the necessary funding. However, this practice is in direct violation of revenue recognition rules. Analyze this situation and determine whether the company's actions constitute blatant fraud. Be sure to discuss the technical and ethical implications of such actions. (4) Evaluate the definite integral $\int_0^{\pi} \cos^2(x) \sin^7(x) dx$ using an appropriate substitution method and provide a step-by-step solution. (5) Speed of boat in still water is 10kmph. If it travels 24km downstream, 16km upstream in the same amount of time, what is the speed of the stream?
Example questions with short answer
(1) What are the parameters that need to be estimated in the general equation of an ellipse, and how do the variables x and y differ from the constants a and b in this context? Provide a detailed explanation of your answer, including the roles of a, b, x, and y in defining the ellipse. (2) Given a company's historical data on revenues, working capital, and net capital expenditures, is it acceptable to forecast change in working capital / net capex by regressing (linear) historical data on revenues? What are the limitations of this approach, and what alternative methods could be used to improve the accuracy of the forecast? (3) Solve the inequality 4(3w+4) $\geq$ 4(2w+12) using interval notation, and express the solution in set-builder notation. (4) Given an image with shape [1,28,28], what will be the shape of the output of a convolution layer with 10 5x5 kernels (filters) without padding? Assume the image dimensions follow the CHW (Channel, Height, Width) format. (5) A gas bubble, from an explosion under water, oscillates with a period T proportional to $P^a * d^b * E^c$. Where 'P' is the static pressure, 'd' is the density of water, and 'E' is the total energy of the explosion. Find the values of a, b, and c, and explain the physical reasoning behind your answer.
Example questions with long answer
(1) Analyze the impact of errors in data analysis on the validity of research findings, using the example of Reinhart and Rogoff's research on the relationship between debt and economic growth. How do such errors affect the development of economic policies, and what are the implications for the field of economics? (2) Prove that the relation $R = \{(1,2),(1,3),(1,4),(2,1),(2,3),(2,4),(3,1),(3,2),(3,4),(4,1),(4,1),(4,3)\}$ is not transitive. Use the definition of transitivity to show that there exist elements x_0, y_0, z_0 such that $(x_0, y_0) \in R$ and $(y_0, z_0) \in R$ but $(x_0, z_0) \notin R$. (3) Astronomers use a new method to detect asteroids and measure their velocity. The method involves detecting energized electromagnetic waves at two different Earth times and using the relative motion of Earth with respect to the asteroid to calculate the velocity. Suppose the Earth spins about 360 degrees within 24 hours, and the asteroid moves in a straight path with respect to other stellar objects. If the angle between the flat Earth surface at point A0 and the direction of asteroid observable is ?, and the angle between the flat Earth surface at point A and the direction of asteroid observable is ?, derive an expression for the relative velocity of Earth with respect to the asteroid at position A0 and A. Use the relativistic Doppler formula to relate the frequencies of the electromagnetic waves detected at points A0 and A. Assume the velocity of the asteroid does not change within the time interval of two detections, and estimate the value of the asteroid's velocity. (4) Bob, a resident outside the US, has purchased a mobile app subscription from a California-based business for under $200. However, due to a showstopper bug, the app is unusable for its main purpose. Bob has attempted to report the issue to the business's support team without success. Discuss the practicalities of Bob suing the California-based business from abroad, considering the requirements of Small Claims court in California and the potential application of consumer protection laws from Bob's home country. How might Bob's approach differ if he were to pursue litigation in a 'normal' court versus Small Claims court, and what are the implications of using an online store like Apple for the purchase? (5) Describe the differences in flow resistance between laminar and turbulent flows in a tube, and explain how the velocity profile changes in the transition from laminar to turbulent flow. Be sure to include the role of viscosity, density, and eddy viscosity in your answer.

Table 5: Human evaluation of question quality across five datasets.

Dataset	MetaMathQA	OpenMathInstruct-2	NuminaMath	WebInstruct	NaturalReasoning
Quality score	5.09	5.37	5.61	5.92	6.46



(a) Average performance across MATH, GPQA, and MMLU-Pro when using varying sizes of data for training Qwen2.5-7B.

(b) Performance breakdown by benchmark, where the highest accuracy per data size is bolded.

SFT Dataset	500K	1M	1.5M
MATH			
WebInstruct	38.82	40.12	38.49
NuminaMath	62.68	–	–
OpenMathInstruct-2	64.25	65.51	65.55
NaturalReasoning	**65.85**	**66.26**	**66.67**
GPQA			
WebInstruct	**31.85**	**30.28**	**29.76**
NuminaMath	1.71	–	–
OpenMathInstruct-2	18.23	14.21	9.08
NaturalReasoning	16.44	20.76	21.35
MMLU-Pro			
WebInstruct	43.65	42.77	43.18
NuminaMath	51.18	–	–
OpenMathInstruct-2	48.96	48.88	49.36
NaturalReasoning	**51.37**	**52.96**	**53.10**

Figure 6: Scaling results for Qwen2.5-7B model.

E Qwen2.5-7B Scaling Results

The scaling trends for the Qwen2.5-7B model plotted by averaging performance on the three benchmarks are shown in Figure 6a. Figure 6b provides a detailed breakdown of model performance across different dataset sizes and benchmarks. It is clear that NATURALREASONING shows superior scaling trends than other datasets.

F Reference Answer Usefulness

F.1 Data Filtering In Knowledge Distillation

We demonstrate the potential usefulness of reference answers using questions from `SelfTrain-15k`. We remove the questions that we are not able to extract a reference answer for and conduct a comparison to understand the utility of reference answers. We fine-tune the Llama3.1-8B-Instruct model using data filtered by final answer verification against a model trained on the unfiltered data.

For final answer verification, we use the prompt in Appendix Figure 10 that prompts the model to judge whether the generated response using Llama3.3-70B-Instruct is in line with the reference final answer, using CoT reasoning. For training data filtering, we only keep the responses that have received a "Yes" final judgement. The training setup includes a learning rate of $1e^{-6}$, a batch size of 64, and training for three epochs, with checkpoints saved every 100 steps for the unfiltered experiment and 50 steps for the filtered experiment due to much smaller data size.

The results are shown in Table 6. Filtering training data using reference answers leads to better performance despite a smaller training set. The filtered dataset contains 7,646 examples, significantly fewer than the 12,349 examples in the unfiltered dataset, yet achieves a higher score on both GPQA-Diamond (32.15 vs. 31.82) and MMLU-Pro (50.06 vs. 49.92). This suggests that higher-quality training data outweighs raw data quantity.

https://doi.org/10.52202/085713-2246

Table 6: SFT results using reference answer filtering. We fine-tune Llama3.1-8B-Instruct on both an unfiltered subsample of NATURALREASONING and a filtered version, where questions are excluded if Llama3.3-70B-Instruct's response disagrees with the reference answer. Despite its smaller size, the filtered set performs better, highlighting the quality of our reference answers.

	Training size	GPQA-Diamond	MMLU-Pro	Average
Llama3.1-8B-Instuct	–	31.82	49.79	40.81
Unfiltered SFT	12,349	31.82	49.92	40.87
Filtered SFT	7,646	**32.15**	**50.06**	**41.11**

Table 7: Online RL results using verifiable rewards on NATURALREASONING. We apply GRPO to Llama3.1-8B-Instruct using the General Verifier as the reward model.

Model	MATH	MMLU-Pro	GPQA	Average
Llama3.1-8B-Instuct	48.74	**49.79**	29.24	42.59
GRPO (Step 50)	**52.20**	49.48	**31.47**	**44.39**

F.2 Reinforcement Learning With Verifiable Rewards

We conduct preliminary experiments applying reinforcement learning with verifiable rewards (RLVR) on the NATURALREASONING dataset. Specifically, we sample a subset of questions whose reference answers are shorter than 10 words and train Llama3.1-8B-Instuct using GRPO [Shao et al., 2024] with the General Verifier [Ma et al., 2025] as the reward model. Training is performed with a batch size of 768. Despite only 50 optimization steps, the model already exhibit noticeable performance gains over the untrained baseline across multiple reasoning benchmarks as shown in Table 7.

These early results suggest that reference answers in NATURALREASONING can be used in RLVR to further enhance reasoning performance, even with limited training. Due to current constraints in computational resources, we leave a more comprehensive exploration of RL-based fine-tuning methods for future work.

G Evaluating NaturalReasoning Against OpenThoughts

To ensure a fair comparison, we adopt a similar knowledge distillation setup, using a stronger reasoning model (i.e., DeepSeek-R1) as the teacher and Llama-3.1-8B-Instruct as the student. Following the procedure in the OpenThoughts paper [Guha et al., 2025], we apply length filtering and then sample 100K questions from each dataset. As shown in Table 8, the model trained on NATURALREASONING outperforms the one trained on OpenThoughts-114k on three out of four benchmarks.

These results show that NATURALREASONING achieves stronger performance on general reasoning and science benchmarks such as GPQA-Diamond, MMLU-Pro, and SuperGPQA [Team et al., 2025], indicating that it complements existing reasoning datasets like OpenThoughts by providing broader coverage and better generalization.

H Cross-Domain Generalization of NaturalReasoning

While our main experiments focus on reasoning benchmarks, we also conduct preliminary studies to examine NATURALREASONING 's applicability to broader knowledge domains. Specifically, we train Llama-3.1-8B-Instruct on 100K randomly sampled NATURALREASONING examples, using DeepSeek-R1 as the teacher model for knowledge distillation. The resulting model is then evaluated on SuperGPQA [Team et al., 2025], which includes diverse subjects beyond mathematics and science—such as philosophy, law, history, economics, literature, and sociology. The results on non-reasoning categories are shown in Table 9.

Despite its focus on reasoning-centric tasks, NATURALREASONING improves performance across a wide range of non-reasoning domains. These results highlight its potential as a versatile foundation for training models with broad domain generalization.

Table 8: Knowledge distillation results comparing training on NATURALREASONING and OpenThoughts. We distill from DeepSeek-R1 into Llama3.1-8B-Instruct using 100K samples from each dataset after applying length filtering following Guha et al. [2025].

Train Set	GPQA-Diamond	MATH-500	MMLU-Pro	SuperGPQA
OpenThoughts-114k (100K)	37.8	**82.2**	59.0	30.0
NaturalReasoning (100K)	**43.1**	69.0	**61.2**	**32.2**

Table 9: Cross-domain evaluation results after training on NATURALREASONING. We fine-tune Llama3.1-8B-Instruct on 100K NATURALREASONING examples with DeepSeek-R1 as the teacher model and evaluate on SuperGPQA's non-reasoning subjects.

Model	Philosophy	Economics	History	Education	Law	Management	Literature	Sociology
Baseline	29.36	23.17	17.24	26.03	**29.63**	29.54	23.37	23.78
Train on NR	**32.56**	**34.86**	**27.49**	**30.17**	**29.63**	**31.14**	**27.08**	**33.57**

I Prompts

The prompt we used for annotating reasoning from the document is shown in Figure 7, Figure 8. We additionally provide the prompt for annotating question validity and difficulty (Figure 9), the prompt used to check if a generated response matches the reference (Figure 10), and the prompt for self scoring (Figure 11).

https://doi.org/10.52202/085713-2246

```
Evaluate the text below according to the scoring instruction and criteria. If the scores are
     high on each axis, derive an exam question following the instructions.

## Scoring Instruction
1. Evaluate the text on each criteria step by step. Provide your honest answer to each sub-
     question. If the answer to a sub-question is a confident Yes, add or subtract the points
     corresponding to the criteria.
2. Keep track of the running points from each criteria to get the total score.
3. Summarize your final evaluation results in a valid JSON object following the instruction
     below.

### Scoring Criteria

**Criteria 1: Problem Completeness**
* The content does not have clear main question, or enough clues to derive the correct answer.
     (0 point)
* The content includes a main question, and enough clues to derive the correct answer. (+1
     point)
* The text shows evidence of engagement and discussion among multiple authors, including
     proposing answers, evaluating and reflecting on answers, responding to critiques, revising
      and editing answers. (+1 point)

**Criteria 2: Problem Complexity and Technical Depth**

* The difficulty of the content is college-level or below. (0 point)
* The difficulty of the content is graduate-level or above, and only domain experts can
     understand. (+1 point)
* The question being discussed is so challenging that even highly skilled non-experts would not
      be able to fully understand the question or provide a correct answer, even after spending
      30 minutes searching the internet or reading up literature. (+1 point)

**Criteria 3: Technical Correctness and Accuracy**

* The text contains significant technical errors or inaccuracies. (-1 point)
* The text demonstrates some technical correctness, but with notable flaws or omissions (e.g.,
     incorrect units, incomplete derivations). (0 point)
* The text demonstrates technical correctness, with some minor flaws or omissions (e.g., minor
     algebraic errors, incomplete explanations). (+0.5 point)
* The text demonstrates high technical correctness, with clear and accurate explanations (e.g.,
      precise definitions, complete derivations). (+0.5 point)
* The text exemplifies exceptional technical correctness, with rigorous and precise
     explanations (e.g., formal proofs, precise calculations). (+1 point)

**Criteria 4: Thinking and Reasoning**

* The text lacks any evidence of thinking or reasoning. (-1 point)
* The text demonstrates some basic thinking and reasoning (+0.5 point), such as:
        + A straightforward application of a known technique.
        + A simple analysis of a problem.
* The text demonstrates some thinking and reasoning (+0.5 point), such as:
        + A consideration of multiple approaches to a problem.
        + A discussion of the trade-offs between different solutions.
* The text demonstrates significant thinking and reasoning (+1 points), such as:
        + Multi-step reasoning chains to solve a complex problem.
        + Advanced reasoning patterns often used in specialized science domains.
* The text exemplifies exceptional thinking and reasoning (+1 points), such as:
        + A highly innovative and creative approach to solving a complex problem in specialized
              domains.
        + Combining multiple reasoning and thinking techniques, with novel abstraction of the
             problem.

## Instruction on Exam Question and Final Report
* Step 1. If BOTH Criteria 1 and Criteria 2 scores above zero, transform the original question
     being discussed to an exam question. The question should focus on problem-solving and
     there should exist a correct answer to the question. The question should be descriptive, i
     .e. use the details and notations from the original text as much as possible. The question
      must be self-contained, concrete, well-defined, i.e. it should NOT contain any missing
     information nor should it contain any ambiguity or subjectiveness.
* Step 2. If BOTH Criteria 1 and Criteria 3 scores above zero, determine whether the text
     contains a correct solution to the question or not. If the discussion DOES contain a
     correct solution, try to extract the gists and important details from the correct answer.
     Then use those key information to derive a correct answer to the question. If there is a
     single final answer, conclude the correct answer with: "Therefore, the final answer is: \\
     boxed{{answer}}.", where [answer] is the number or expression that is the final answer.
* Step 3. If BOTH Criteria 2 and Criteria 4 have non-zero scores, write down a list of critical
      knowledge and reasoning steps which are required to derive a correct answer to the exam
     question. Each item in the list must be descriptive, specific and concrete.
* Step 4. Label the question difficulty with Easy, Medium, Hard, and Extra Hard.
```

Figure 7: Prompt for annotating reasoning from the document, generating a question and reference answer. (Part 1)

```
Finally, copy all your analysis from above into a JSON object at the end of the final report.
     The JSON object should contain the following attributes:
- ‘"scores"‘: a list of dictionary entries, where each entry contains the criteria name and
     corresponding score on that criteria.
- ‘"exam\_question"‘: a string recording the full exam question derived from Step 1 analysis.
     DO NOT omit any details. If the scores for Criteria 1 and Criteria 2 are low and no exam
     question can be made out of the text, return an empty string.
- ‘"correct\_answer"‘: a string recording the correct answer derived from Step 2. If the
     discussion does not contain a correct answer, return an empty string.
- ‘"knowledge\_and\_reasoning\_steps"‘: a list of strings, where each entry copying the
     critical piece of knowledge or important reasoning steps derived from Step 3. If either
     Criteria 2 or Criteria 4 has score zero, return an empty list.
- ‘"question\_difficulty"‘: a string recording the difficulty of the question derived from Step
      4.

### Text
{text}
```

Figure 8: Prompt for annotating reasoning from the document, generating a question and reference answer. (Part 2)

```
Your task is to verify and improve the quality of a question.

A valid question must meet the following criteria:
* The question should contain a problem to be solved, instead of
     only presenting statements.
* The question should be well-defined and self-contained, i.e.
    have all the necessary information to derive an answer.
* The question should be specific and clear. There should be one
     correct answer to the question.
* The question should not refer to external resources, such as
    figures, videos, etc.
* To derive an answer, multi-step reasoning and recalling
    relevant knowledge is required.
* The difficulty should be graduate-level or above.
* The question can contain LaTex but it should be correct.
If a question does not meet any of the criterion above, revise
    it till it meets all the criteria.

IMPORTANT: Put your final answer in a JSON object with two
    fields:
- "question_quality_score": rate how well the question meets all
     the criteria, on a scale of 1 to 10.
- "improved_question": revised question which will meet the
    criteria better and thus has a higher question quality score.

Question:
{question}
```

Figure 9: Prompt for annotating quality scores.

https://doi.org/10.52202/085713-2246

```
You are an expert evaluator tasked with deciding whether a
    response meets the standards of quality and correctness for
    general reasoning tasks. Your evaluation must consider both
    the quality of the reasoning process (chain of thought, CoT)
     and the correctness or appropriateness of the final answer.

### Evaluation Criteria:
1. **Correctness of the Final Answer**:
   - Does the final answer align with the reference answer or
       the expected outcome?
2. **Quality of the Thinking Process (CoT)**:
   - Is the reasoning logical, coherent, and free from
       significant errors?
   - Does the reasoning support the final answer in a clear and
        step-by-step manner?
3. **Completeness**:
   - Does the response adequately address all aspects of the
       instruction or problem?

### Input Details:
- **Instruction**: {Describe the task, problem, or instruction
    here.}
- **Reference Answer**: {Provide the expected or ideal outcome
    here.}
- **Response**: {Include the response to be evaluated,
    containing both the CoT and the final answer.}

### Task:
Analyze the response provided and decide if it is a "Yes" or "
    No" based on the following:
- **"Yes"**: The response meets the required standards for
    correctness, reasoning, and completeness.
- **"No"**: The response fails to meet one or more of the
    standards.

Provide a brief explanation of your decision, highlighting
    specific strengths or weaknesses in the reasoning process (
    CoT), the final answer, or completeness.

**Response Format**:
1. **Explanation**:
   - **Final Answer Evaluation**: (Discuss correctness and
       consistency.)
   - **Chain of Thought Evaluation**: (Discuss logic and
       coherence.)
   - **Completeness**: (Assess whether the response fully
       addresses the instruction.)
2. **Judgment**: Yes/No
---

[FEW-SHOT EXAMPLES HERE]
---

### Current Input:
**Instruction**: <INSTRUCTION>
**Reference Answer**: <ANSWER>
**Response**:
<RESPONSE>

**Evaluation**:
```

Figure 10: Prompt used to check if a response matches the reference answer.

```
You are an expert evaluator tasked with analyzing a response to
    a general reasoning problem. Your goal is to determine if the
     response demonstrates good reasoning (CoT) and whether the
    reasoning makes sense overall.

### **Evaluation Criteria**:
1. **Reasoning Quality (CoT)**:
   - Does the reasoning follow a logical and coherent sequence?
   - Are the steps valid and free of major errors?
   - Does the reasoning align with standard problem-solving
       practices?

2. **Accuracy**:
   - Does the reasoning lead to the correct or valid conclusion
       based on the instruction?

3. **Clarity**:
   - Is the response clear and easy to understand?

### Input Details:
- **Instruction**: {Describe the task, problem, or instruction
    here.}
- **Response**: {Include the response to be evaluated,
    containing the chain of thought (CoT).}

### **Task**:
Analyze the response and decide if it meets the standards for
    correctness, reasoning quality, and clarity. Provide your
    judgment as either **"Yes"** (the response is good) or **"No
    "** (the response is not good). Then, briefly explain your
    decision.

**Response Format**:
1. **Judgment**: Yes/No
2. **Explanation**:
   - **Reasoning Quality (CoT)**: (Assess the reasoning process
       in detail.)
   - **Accuracy**: (Evaluate whether the reasoning leads to the
       correct conclusion.)
   - **Clarity**: (Comment on the clarity of the response.)

---
[FEW-SHOT EXAMPLES]

---

### Current Input:
**Instruction**: <INSTRUCTION>
**Response**:
<RESPONSE>

**Evaluation**:
1. **Judgment**:
```

Figure 11: Prompt for self scoring.

https://doi.org/10.52202/085713-2246

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We present a new dataset containing 2.8M challenging reasoning questions and demonstrate its effectiveness in training reasoning models through knowledge distillation and self-training experiments.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: We have a Limitation Section.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This is an empirical paper driven by experiments.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have included our training and evaluation details in the experiment sections.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

 https://doi.org/10.52202/085713-2246

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We release the 2.8M questions we collected to the community to foster research discoveries. While we do not release code for our experiments, all the expeirment details are documented in the paper which should be easy to reproduce.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The training and evaluation details are well documented in experiment sections

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: For all the evaluation results, we average the performance of last three checkpoint to reduce variance.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: We used different types of compute resources for different experiments, it was hard to enumerate the resources we used.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

https://doi.org/10.52202/085713-2246

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have a Impact Statement section

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have properly cited the prior works.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We release a new dataset along with this paper, and we have well documented the dataset on HuggingFace.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

https://doi.org/10.52202/085713-2246

Answer: [Yes]

Justification: Yes, we have included details about human evaluation of datasets in Appendix D.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: There is no such risk in our study.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.