NaturalReasoning: Reasoning in the Wild with 2.8M Challenging Questions

Weizhe Yuan, Jane Yu, Song Jiang, Karthik Padthe, Yang Li, Ilia Kulikov, Kyunghyun Cho, Dong Wang, Yuandong Tian, Jason Weston, Xian Li

Motivation

- Recent LLMs are getting much better at reasoning, thanks to the advancement in test-time scaling (i.e. letting the model “think more”)

Motivation

- However...
  - Current methods for training reasoning models rely heavily on Reinforcement Learning with Verifiable Rewards

Motivation

- However...
  - Current methods for training reasoning models rely heavily on Reinforcement Learning with Verifiable Rewards
    - Limited tasks: Mainly consider verifiable tasks such as maths/coding

Motivation

- However…
  - Current methods for training reasoning models rely heavily on Reinforcement Learning with Verifiable Rewards
    - Limited tasks: Mainly consider verifiable tasks such as maths/coding
    - LOTS of other reasoning domains (e.g., physics, economics) remain underexplored

Motivation

- However…
  - Current methods for training reasoning models rely heavily on Reinforcement Learning with Verifiable Rewards
    - Limited tasks: Mainly consider verifiable tasks such as maths/coding
    - LOTS of other reasoning domains (e.g., physics, economics) remain underexplored
    - LOTS of other open-ended reasoning tasks remain underexplored

Research Question 🤔

- Can we automatically generate large-scale reasoning tasks that span diverse disciplines and demand extended reasoning to solve?

Our Approach

- **Key idea**: Use an LLM to synthesize challenging questions based on pre-training corpus

Data Collection

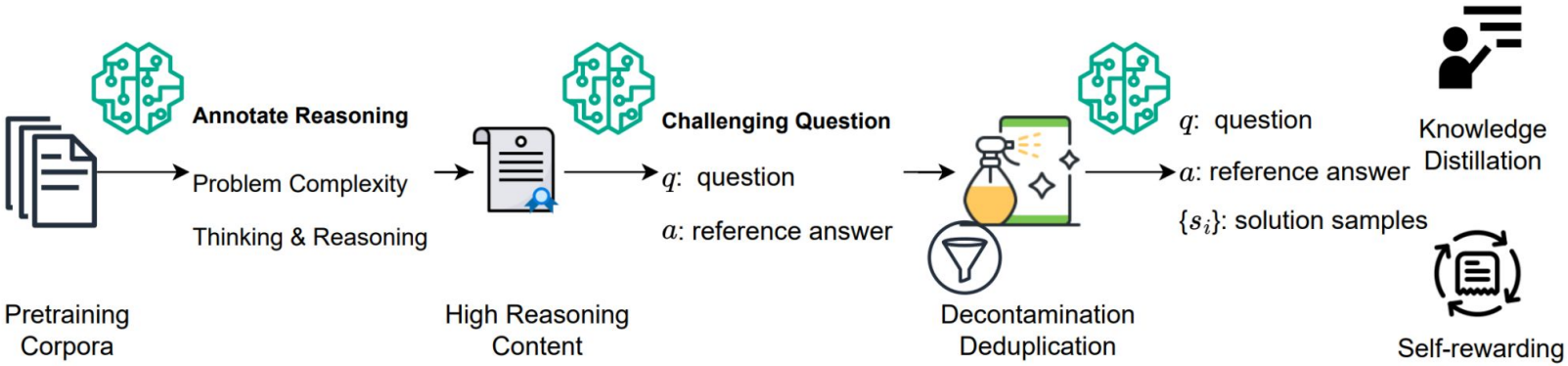


Figure 1: An overview of the data creation approach of NATURALREASONING.

Data Collection

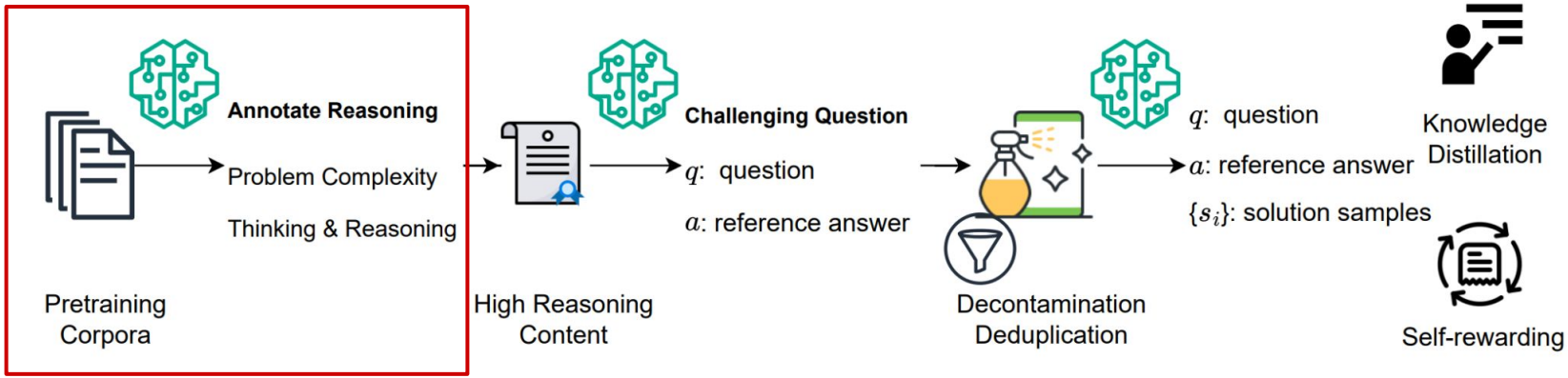


Figure 1: An overview of the data creation approach of NATURALREASONING.

Step 1: Annotate Reasoning

- Given a document, we prompt an LLM to rate the content in the document along multiple axes: Problem Completeness, Technical Depth, Thinking and Reasoning, etc.

Data Collection

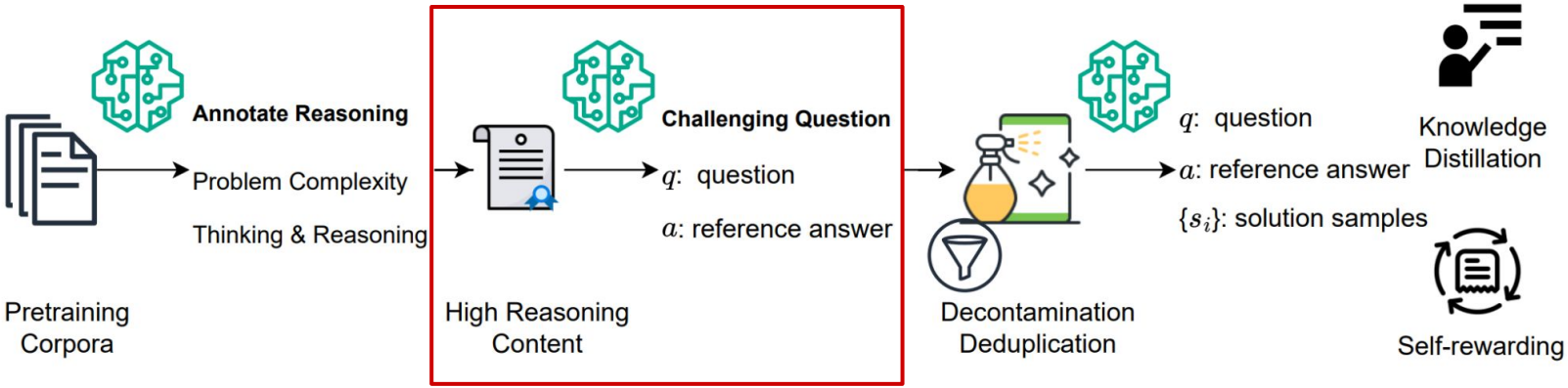


Figure 1: An overview of the data creation approach of NATURALREASONING.

Step 2: Synthesize Questions

- For document that contains high reasoning content, we further prompt an LLM to compose a self-contained and challenging reasoning question with reference answer. For every question we generate an additional response with Llama-3.3-70B-Instruct

Data Collection

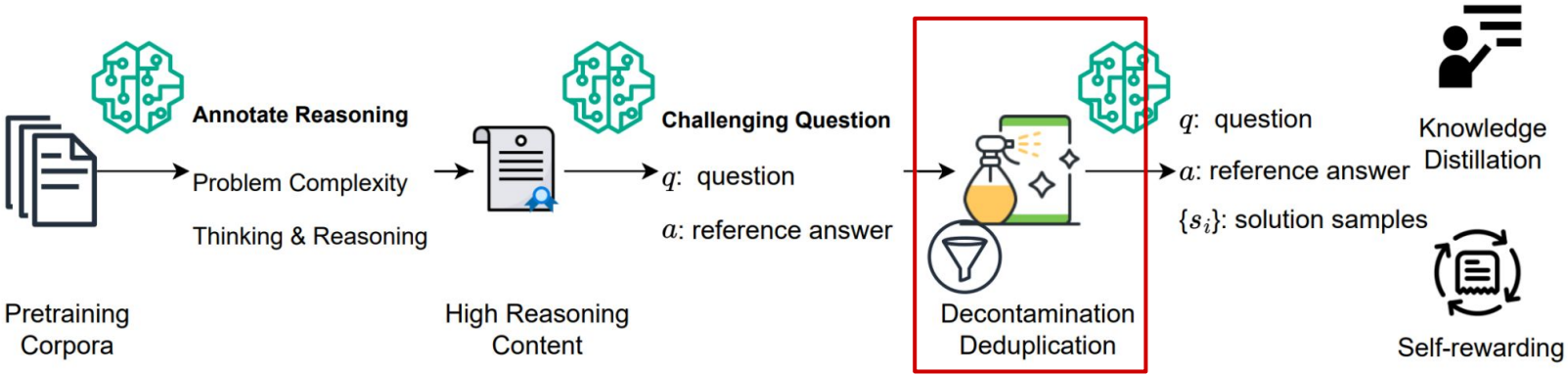


Figure 1: An overview of the data creation approach of NATURALREASONING.

Step 3: Deduplication & Decontamination

Data Collection

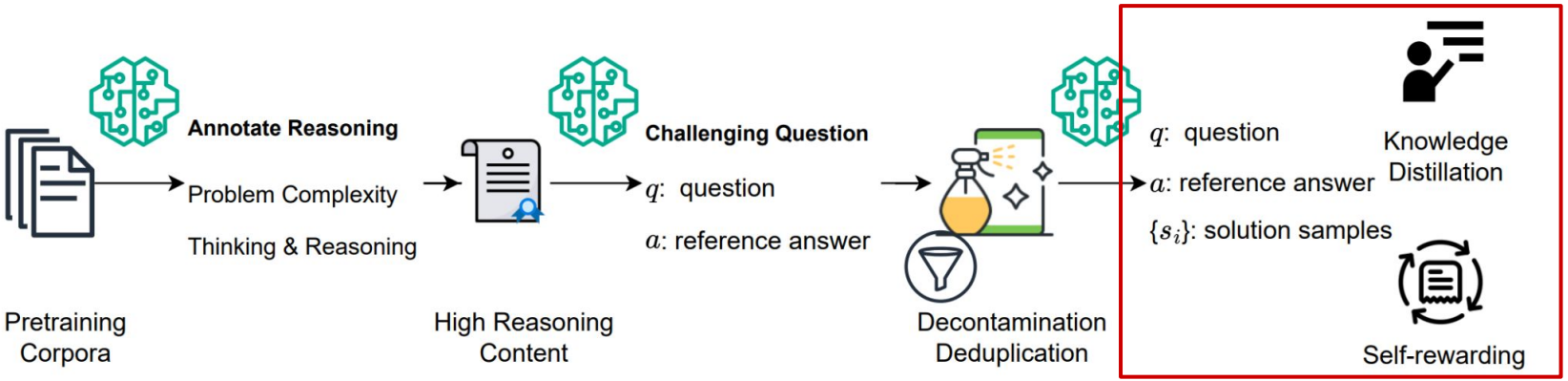


Figure 1: An overview of the data creation approach of NATURALREASONING.

Finally, the questions could be used for training reasoning models through either (1) knowledge distillation from a stronger model, (2) unsupervised self-training using external reward models or self-rewarding

Data Collection

- We collected a total of 2.8M questions, forming the NaturalReasoning dataset.

Eliciting Long Chain-of-Thought through Knowledge Distillation

- Are NaturalReasoning questions hard enough to elicit long CoTs? 🤔

Eliciting Long Chain-of-Thought through Knowledge Distillation

- We start from a Llama-3.3-70B-Instruct model and distill responses from DeepSeek-R1 through Supervised Fine-tuning.
  - We hypothesize that questions from NaturalReasoning are challenging enough to elicit long CoTs
  - **Comparison Question Set**: s1k, LIMO

Eliciting Long Chain-of-Thought through Knowledge Distillation

	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

- Randomly selected 1k NATURALREASONING matches—even slightly exceeds—the performance obtained on datasets that underwent several rounds of meticulous filtering and curation

Eliciting Long Chain-of-Thought through Knowledge Distillation

	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

- Performance increases monotonically as we scale the data size.

Unsupervised Self-Training

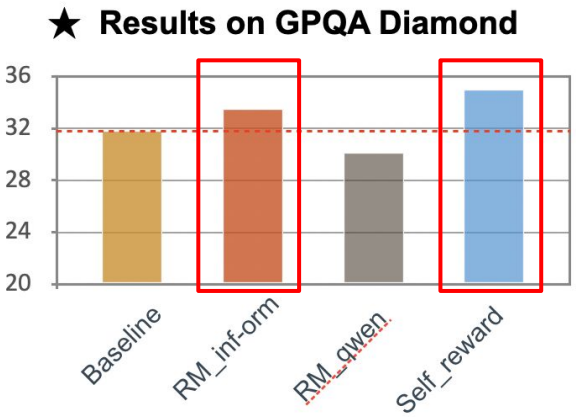
- Can NaturalReasoning questions be used for unsupervised self-training? 🤔

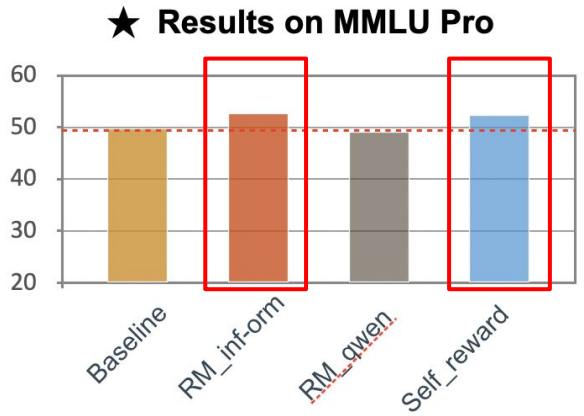
Unsupervised Self-Training

- We start from a Llama3.1-8b (instruct) model, aiming to improve it through self training.
  - **Dataset:** We use a subset of 15k questions from our NaturalReasoning dataset to do self-training loop
  - We conduct one iteration of DPO and compare the performance of using strong external reward models (INF-ORM-Llama3.1-70B, Qwen2.5-Math-RM-72B) and self-reward

Unsupervised Self-Training

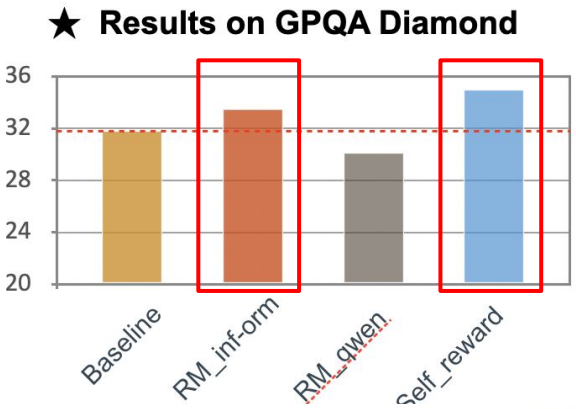
- Reasoning performance on GPQA Diamond and MMLU-Pro
  - **Training on NaturalReasoning questions can improve LLM reasoning with strong external RMs or self-rewarding**

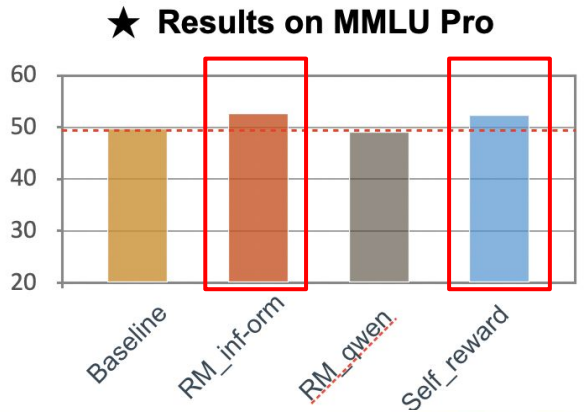




Unsupervised Self-Training

- Reasoning performance on GPQA Diamond and MMLU-Pro
  - **Training on NaturalReasoning questions can improve LLM reasoning with strong external RMs or self-rewarding**





Feel free to check out our paper or stop by our poster session (Wed 3 Dec 4:30 p.m. PST — 7:30 p.m. PST) to see more results and analysis!

Thanks for Listening