General-Reasoner: Advancing LLM Reasoning Across All Domains

Xueguang Ma*, Qian Liu*, Dongfu Jiang, Ge Zhang, Zejun Ma, Wenhu Chen

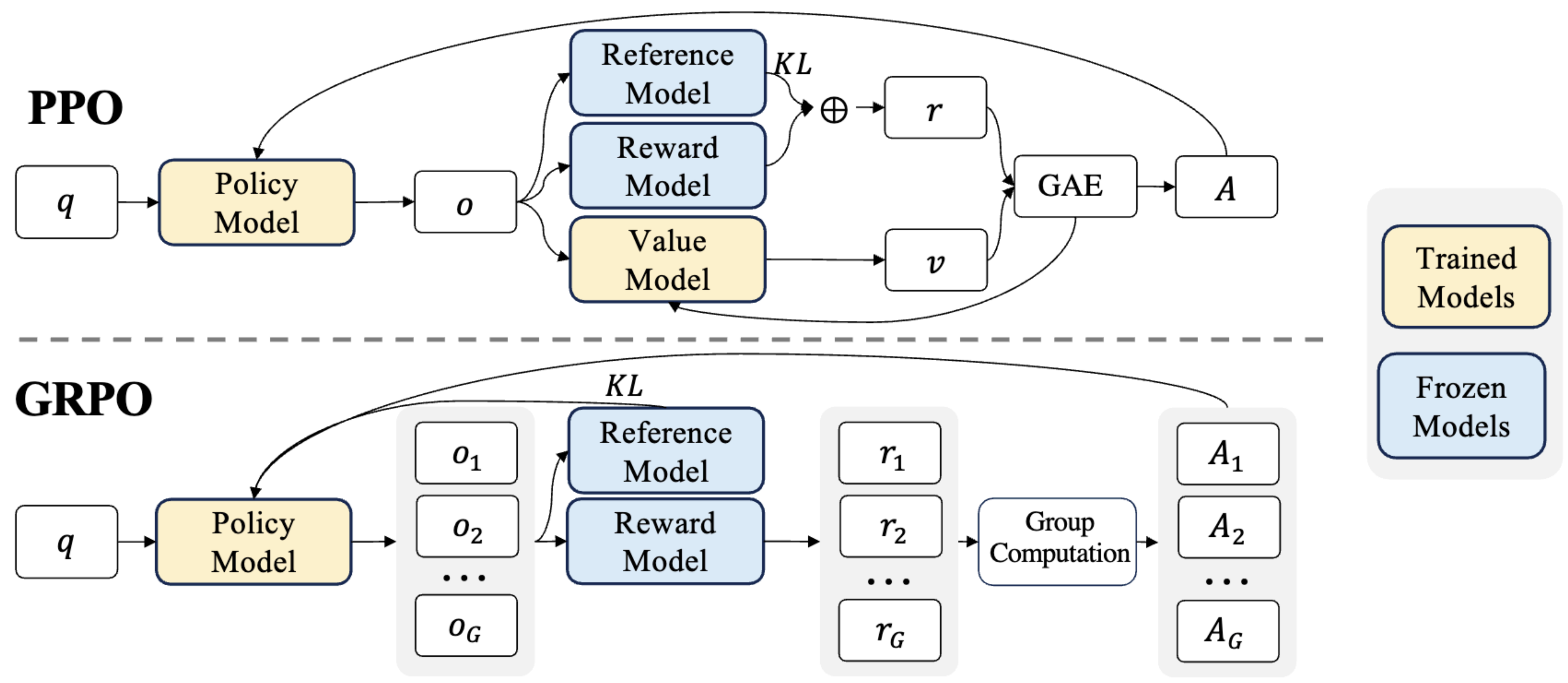
UNIVERSITY OF WATERLOO

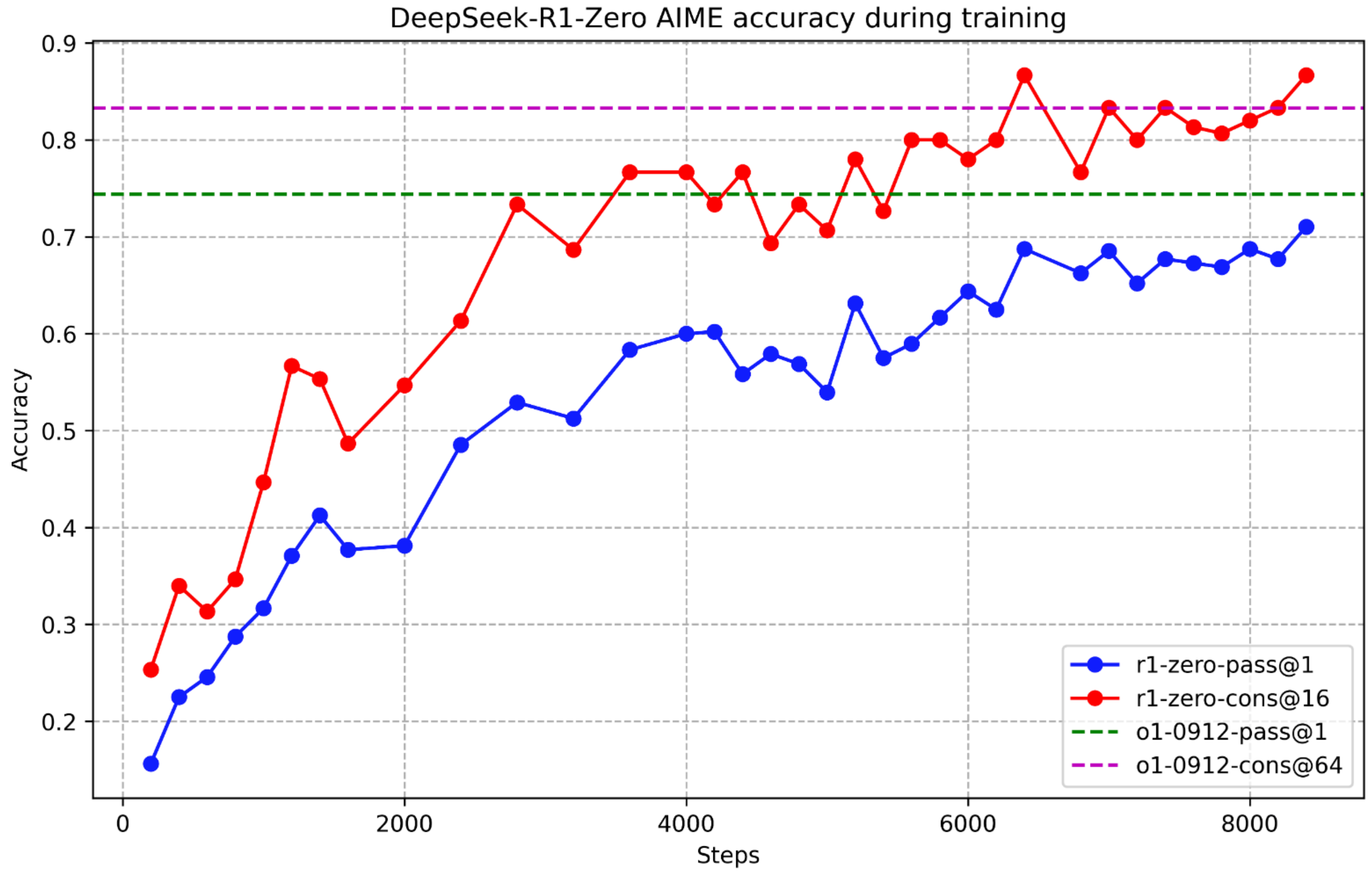
TikTok

NEURAL INFORMATION PROCESSING SYSTEMS

Background: RL for Reasoning Tasks

The success of reinforcement learning with verifiable reward (RLVR) in improving the reasoning capability of large language model (LLM).

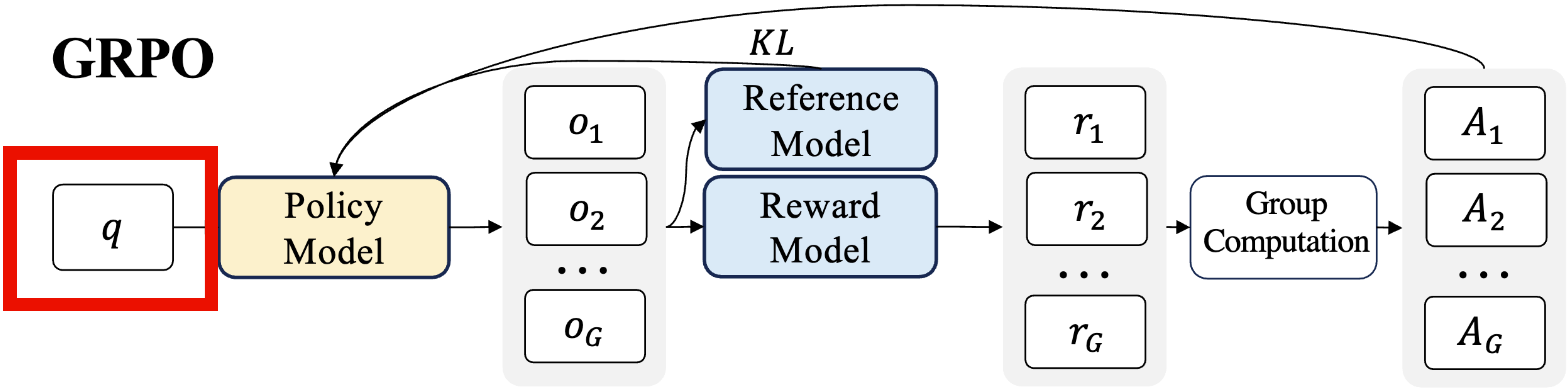




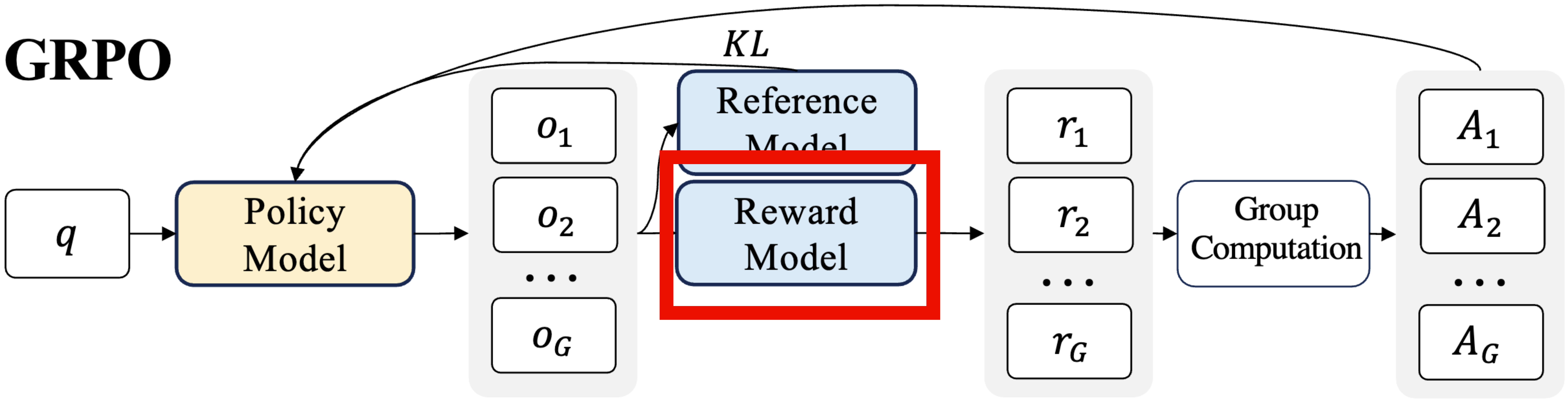
DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models

DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

What Data to Use?



How to Assign Reward?



Limitations of Existing RLVR Training

- Data:
  - Limited to Math domain and Code domain.

- Reward:
  - Computed by rule based verifier like string matching for math, or test cases for code.

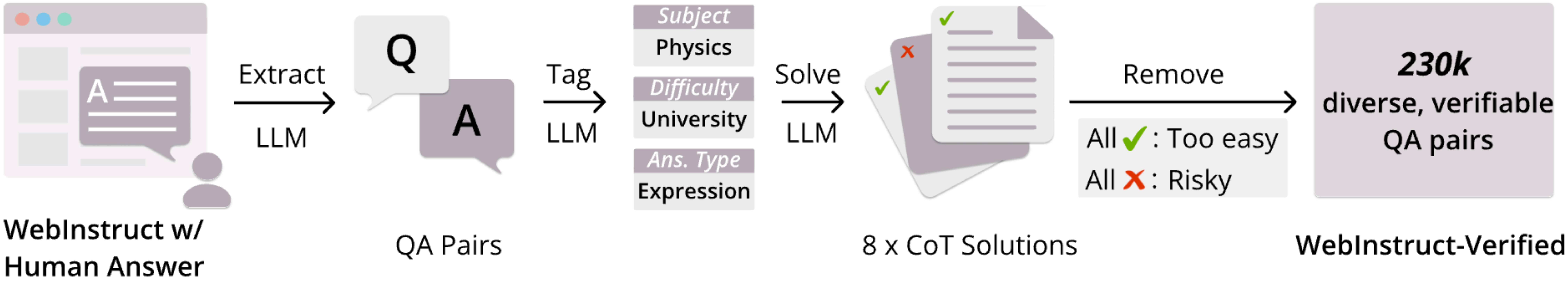
How to effectively scale RLVR for all domain?

General-Reasoner: Advancing LLM Reasoning Across All Domains

- Data:
  - Scale the verifiable reasoning QA data by extracting from web crawled document.

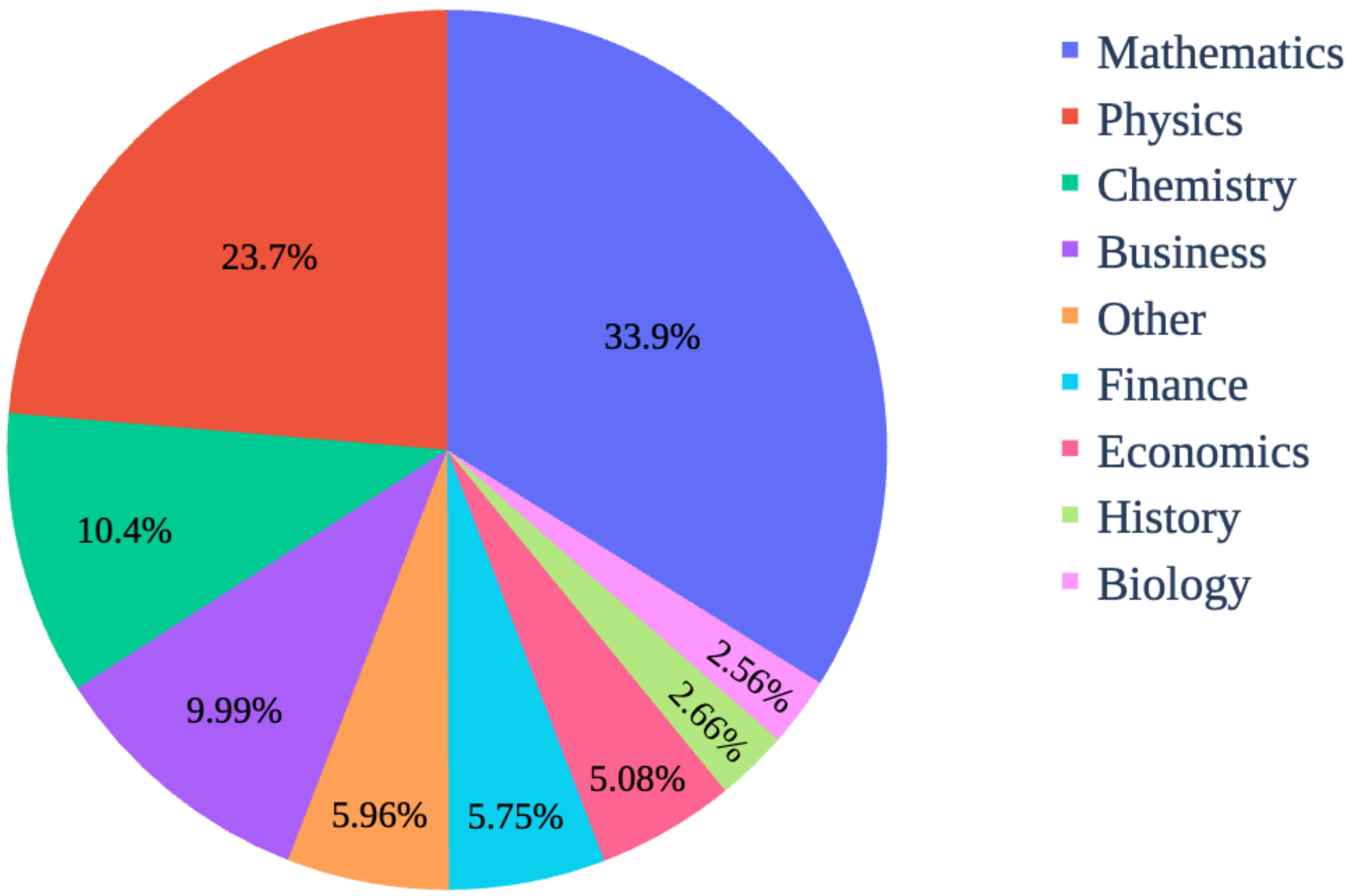
- Reward:
  - To handle more varied answer format, using model-based generative verifier to assign reward score.

Scaling the Data: WebInstruct-Verified

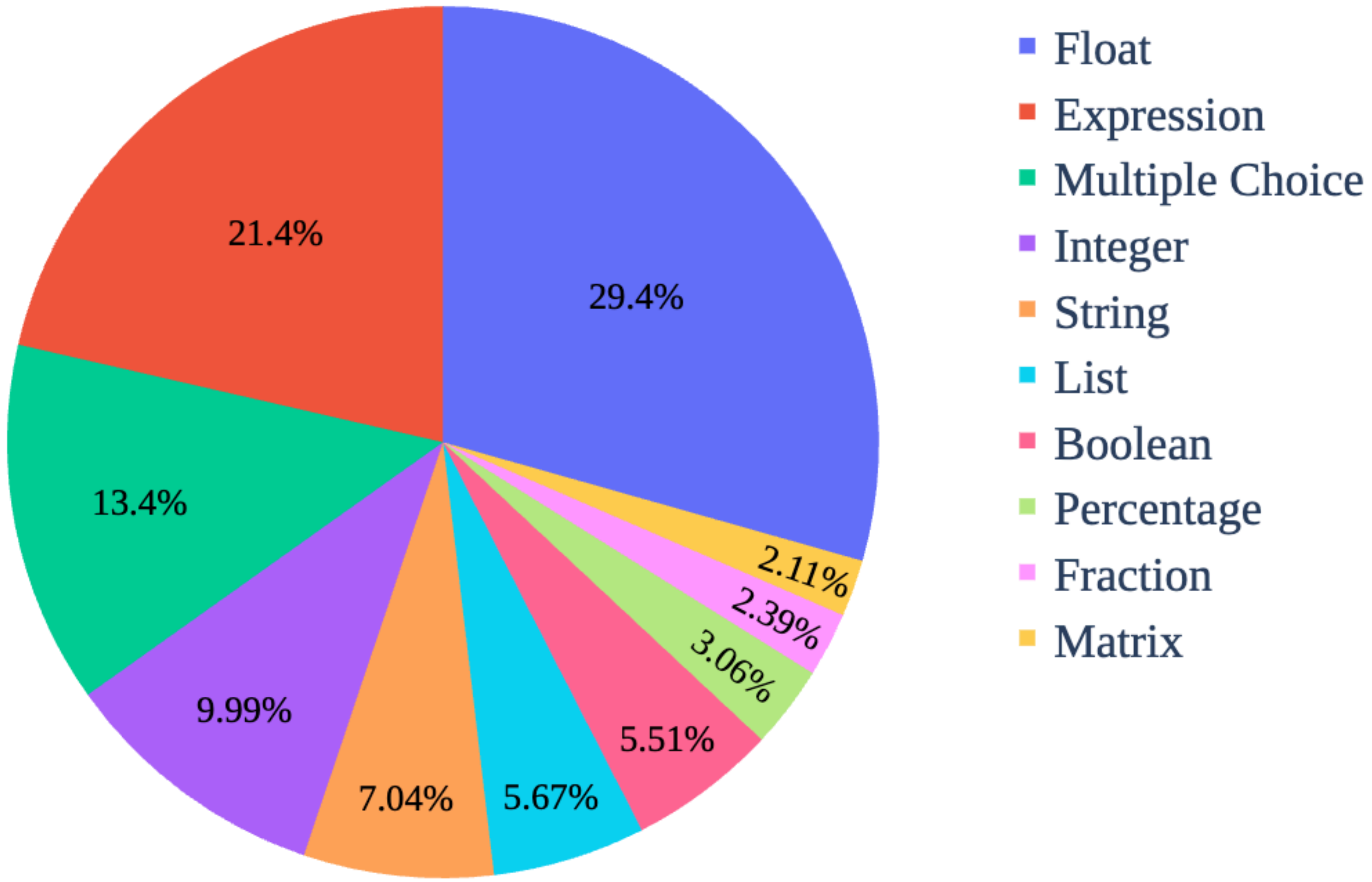


Scaling the Data: WebInstruct-Verified

Domains



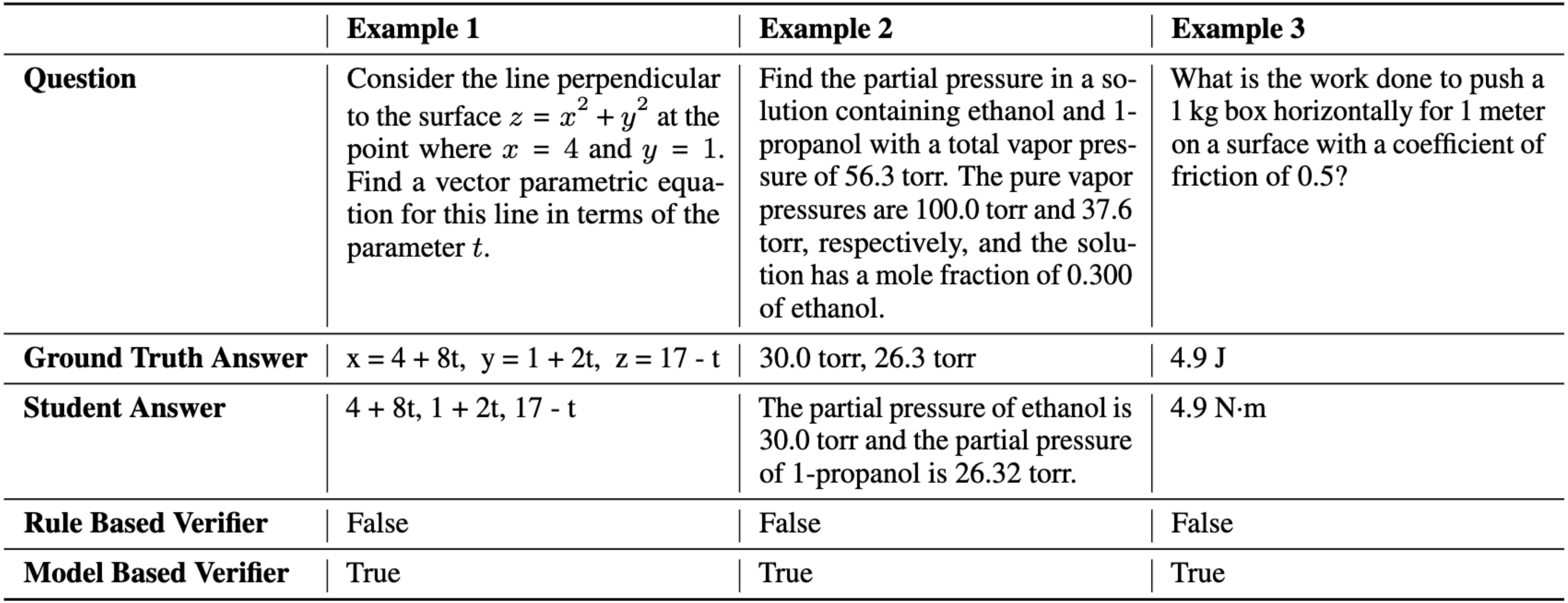
Answer Formats



Improving Reward Assignment: General-Verifier

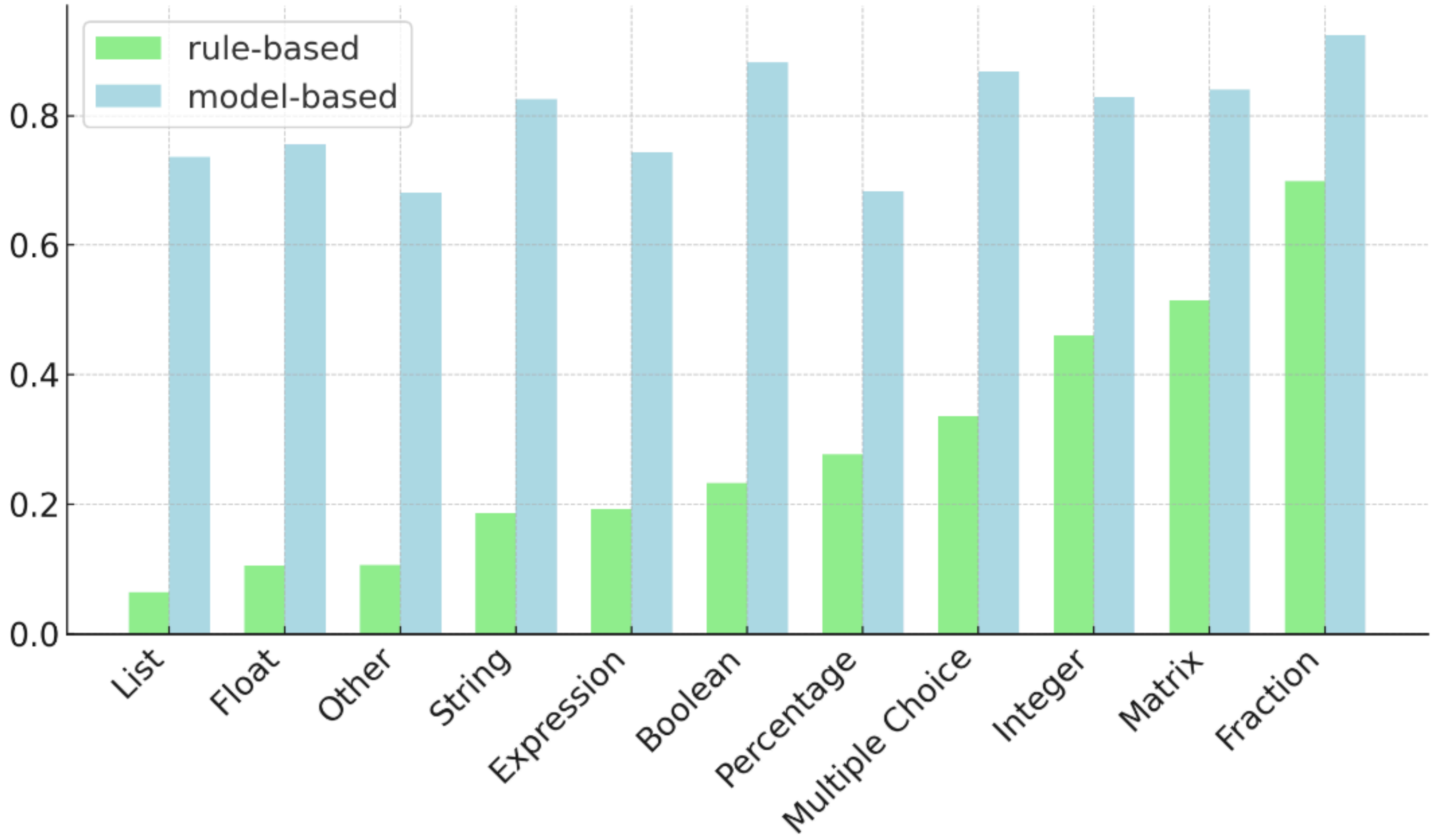
- Given question Q, prompt Gemini-2.0 to generate CoT with short answer A’ concluded.
- Then, use Gemini-2.0 to generate CoT that compares A and A’.
- In this way, synthesize large-scale input-output pairs:
  - (Q, A, A’) → (CoT, V), where V is the verdict (equal or not equal).
- Train a compact model like Qwen2.5-Math-1.5B specifically for answer verification.
  - Effective verification
  - Efficient Inference

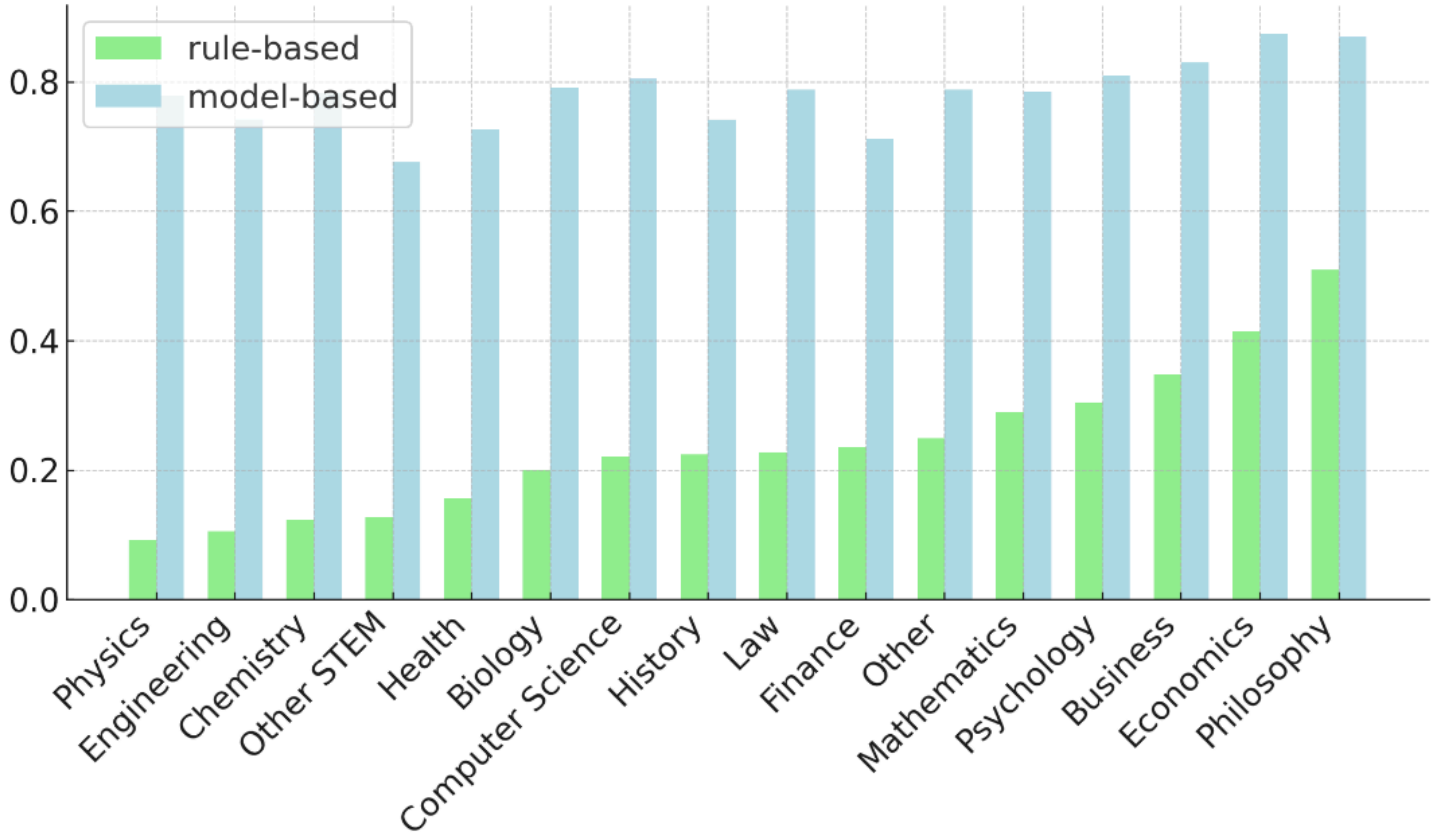
General-Verifier v.s. Rule-Based Verifer



	Example 1	Example 2	Example 3
Question	Consider the line perpendicular to the surface $z = x^2 + y^2$ at the point where $x = 4$ and $y = 1$. Find a vector parametric equation for this line in terms of the parameter t.	Find the partial pressure in a solution containing ethanol and 1-propanol with a total vapor pressure of 56.3 torr. The pure vapor pressures are 100.0 torr and 37.6 torr, respectively, and the solution has a mole fraction of 0.300 of ethanol.	What is the work done to push a 1 kg box horizontally for 1 meter on a surface with a coefficient of friction of 0.5?
Ground Truth Answer	x = 4 + 8t, y = 1 + 2t, z = 17 - t	30.0 torr, 26.3 torr	4.9 J
Student Answer	4 + 8t, 1 + 2t, 17 - t	The partial pressure of ethanol is 30.0 torr and the partial pressure of 1-propanol is 26.32 torr.	4.9 N·m
Rule Based Verifier	False	False	False
Model Based Verifier	True	True	True

General-Verifier v.s. Rule-Based Verifer



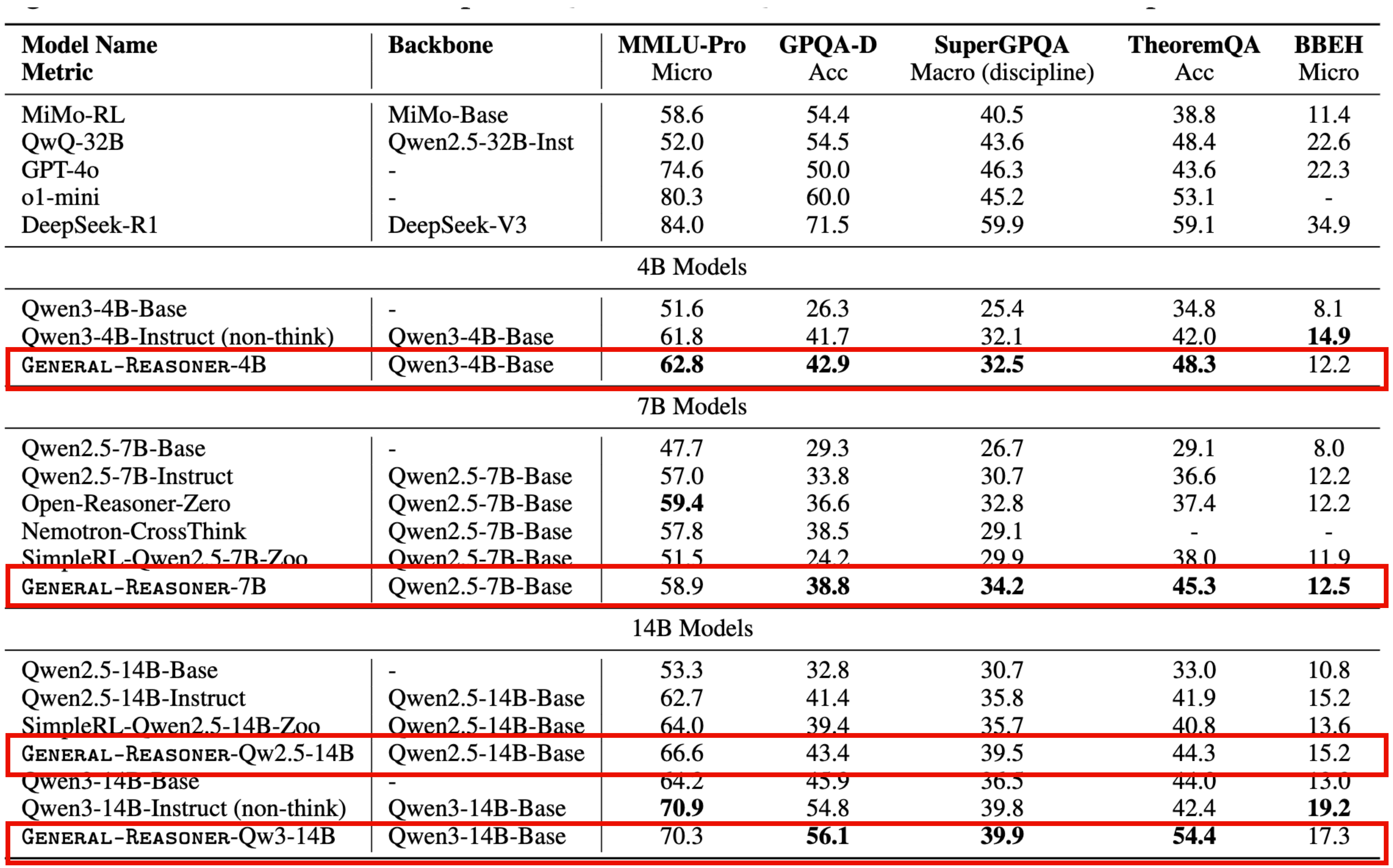


Experiment Setup

- Zero-RL training with GRPO
  - Directly train from Base model, e.g. Qwen3-4B-Base
  - Implemented based on verl

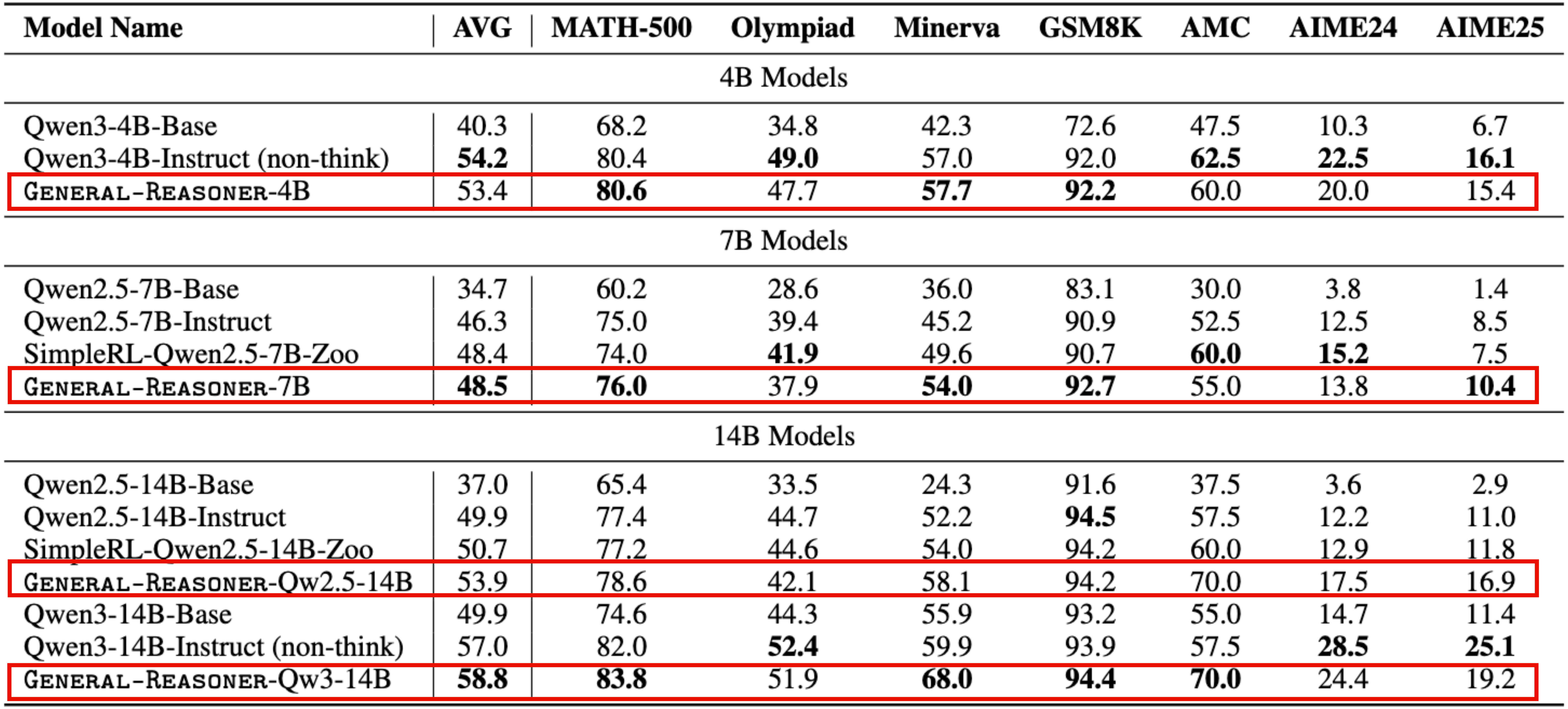
- Comprehensive evaluation on math and general reasoning tasks
  - MMLU-PRO, SuperGPQA, BBEH, TheoremQA etc.

Experiment Results: General Domain Reasoning



Model Name Metric	Backbone	MMLU-Pro Micro	GPQA-D Acc	SuperGPQA Macro (discipline)	TheoremQA Acc	BBEH Micro
MiMo-RL	MiMo-Base	58.6	54.4	40.5	38.8	11.4
QwQ-32B	Qwen2.5-32B-Inst	52.0	54.5	43.6	48.4	22.6
GPT-4o	-	74.6	50.0	46.3	43.6	22.3
o1-mini	-	80.3	60.0	45.2	53.1	-
DeepSeek-R1	DeepSeek-V3	84.0	71.5	59.9	59.1	34.9
4B Models						
Qwen3-4B-Base	-	51.6	26.3	25.4	34.8	8.1
Qwen3-4B-Instruct (non-think)	Qwen3-4B-Base	61.8	41.7	32.1	42.0	**14.9**
General-Reasoner-4B	Qwen3-4B-Base	**62.8**	**42.9**	**32.5**	**48.3**	12.2
7B Models						
Qwen2.5-7B-Base	-	47.7	29.3	26.7	29.1	8.0
Qwen2.5-7B-Instruct	Qwen2.5-7B-Base	57.0	33.8	30.7	36.6	12.2
Open-Reasoner-Zero	Qwen2.5-7B-Base	**59.4**	36.6	32.8	37.4	12.2
Nemotron-CrossThink	Qwen2.5-7B-Base	57.8	38.5	29.1	-	-
SimpleRL-Qwen2.5-7B-Zoo	Qwen2.5-7B-Base	51.5	24.2	29.9	38.0	11.9
General-Reasoner-7B	Qwen2.5-7B-Base	58.9	**38.8**	**34.2**	**45.3**	**12.5**
14B Models						
Qwen2.5-14B-Base	-	53.3	32.8	30.7	33.0	10.8
Qwen2.5-14B-Instruct	Qwen2.5-14B-Base	62.7	41.4	35.8	41.9	15.2
SimpleRL-Qwen2.5-14B-Zoo	Qwen2.5-14B-Base	64.0	39.4	35.7	40.8	13.6
General-Reasoner-Qw2.5-14B	Qwen2.5-14B-Base	66.6	43.4	39.5	44.3	15.2
Qwen3-14B-Base	-	64.2	45.9	36.5	44.0	13.0
Qwen3-14B-Instruct (non-think)	Qwen3-14B-Base	**70.9**	54.8	39.8	42.4	**19.2**
General-Reasoner-Qw3-14B	Qwen3-14B-Base	70.3	**56.1**	**39.9**	**54.4**	17.3

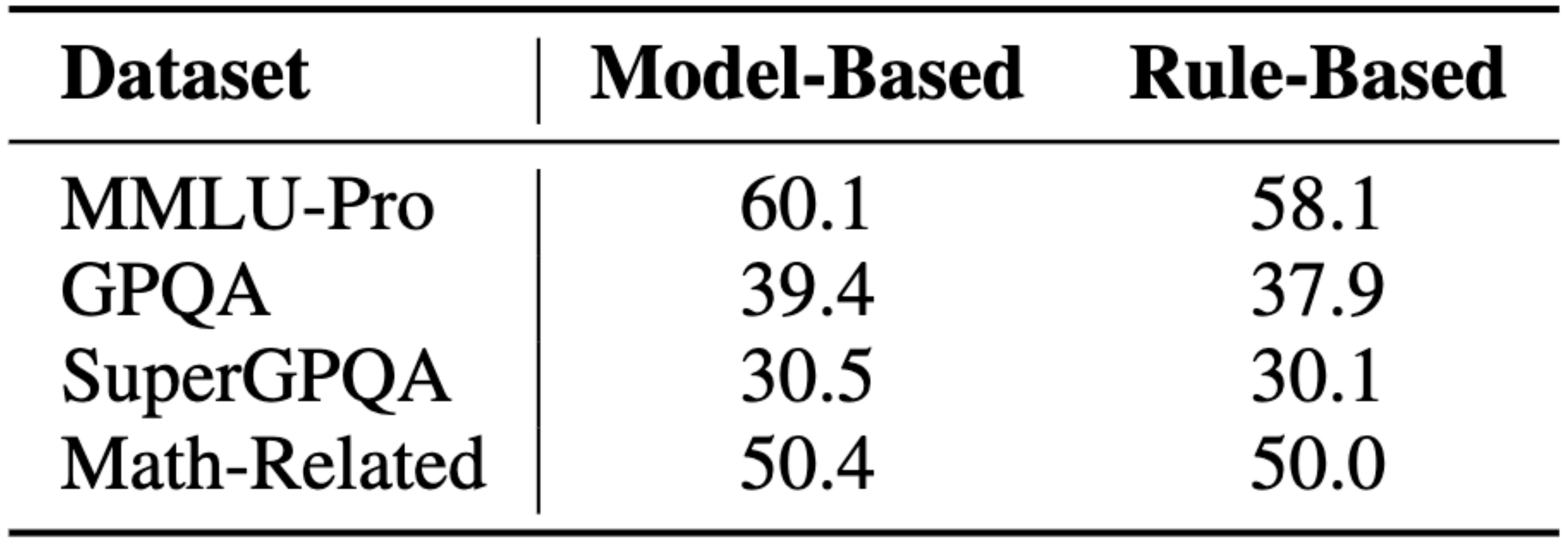
Experiment Results: Math Reasoning



Model Name	AVG	MATH-500	Olympiad	Minerva	GSM8K	AMC	AIME24	AIME25
4B Models								
Qwen3-4B-Base	40.3	68.2	34.8	42.3	72.6	47.5	10.3	6.7
Qwen3-4B-Instruct (non-think)	**54.2**	80.4	**49.0**	57.0	92.0	**62.5**	**22.5**	**16.1**
General-Reasoner-4B	53.4	**80.6**	47.7	**57.7**	**92.2**	60.0	20.0	15.4
7B Models								
Qwen2.5-7B-Base	34.7	60.2	28.6	36.0	83.1	30.0	3.8	1.4
Qwen2.5-7B-Instruct	46.3	75.0	39.4	45.2	90.9	52.5	12.5	8.5
SimpleRL-Qwen2.5-7B-Zoo	48.4	74.0	**41.9**	49.6	90.7	**60.0**	**15.2**	7.5
General-Reasoner-7B	**48.5**	**76.0**	37.9	**54.0**	**92.7**	55.0	13.8	**10.4**
14B Models								
Qwen2.5-14B-Base	37.0	65.4	33.5	24.3	91.6	37.5	3.6	2.9
Qwen2.5-14B-Instruct	49.9	77.4	44.7	52.2	**94.5**	57.5	12.2	11.0
SimpleRL-Qwen2.5-14B-Zoo	50.7	77.2	44.6	54.0	94.2	60.0	12.9	11.8
General-Reasoner-Qw2.5-14B	53.9	78.6	42.1	58.1	94.2	70.0	17.5	16.9
Qwen3-14B-Base	49.9	74.6	44.3	55.9	93.2	55.0	14.7	11.4
Qwen3-14B-Instruct (non-think)	57.0	82.0	**52.4**	59.9	93.9	57.5	**28.5**	**25.1**
General-Reasoner-Qw3-14B	**58.8**	**83.8**	51.9	**68.0**	**94.4**	**70.0**	24.4	19.2

Ablation: Effectiveness of General Verifier

Table 5: Zero RL training using our model-based verifier versus the rule-based verifier on the Qwen3-4B-Base model for 120 step.



Dataset	Model-Based	Rule-Based
MMLU-Pro	60.1	58.1
GPQA	39.4	37.9
SuperGPQA	30.5	30.1
Math-Related	50.4	50.0

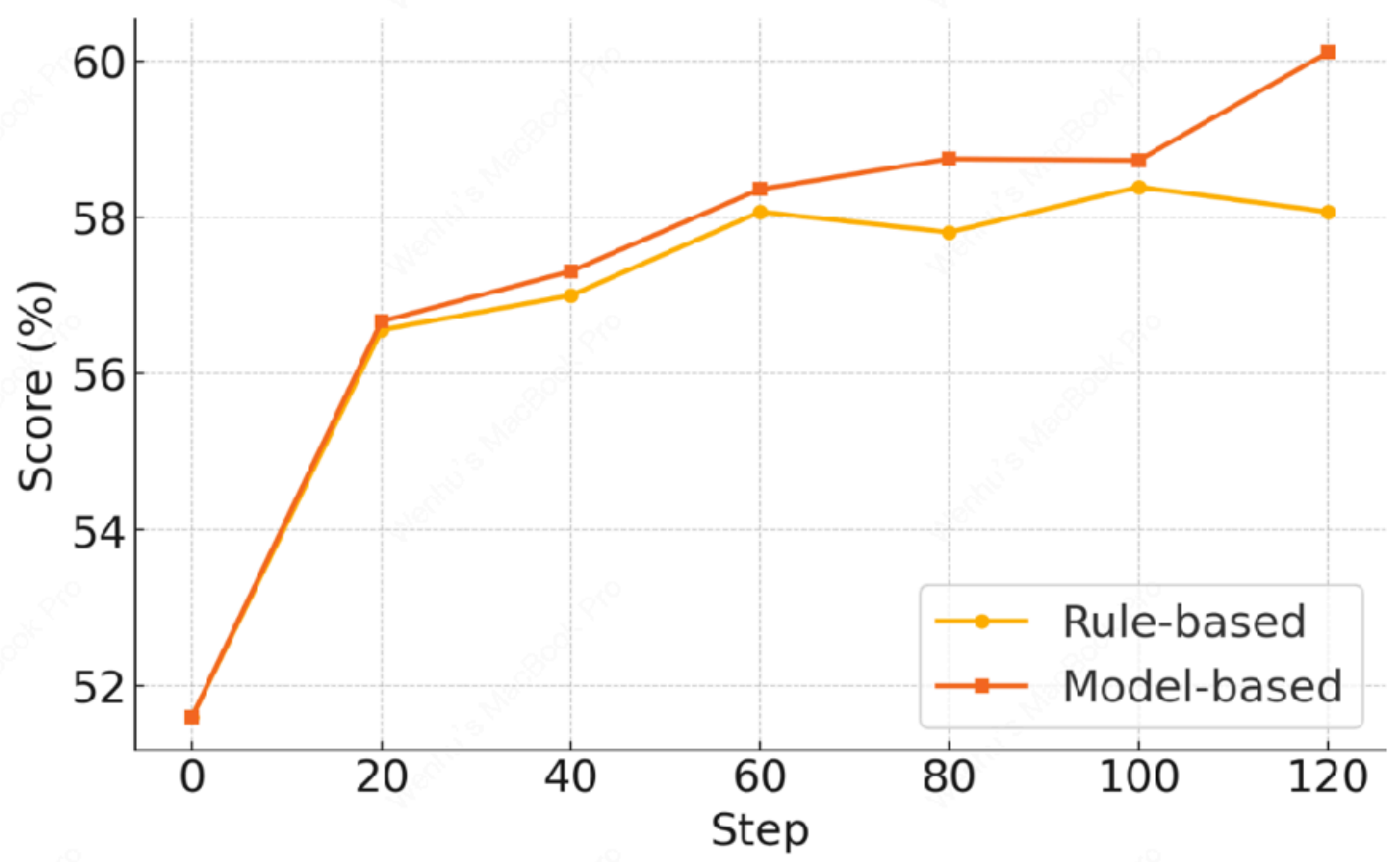


Figure 4: MMLU-Pro evaluation score at different training step using model-based verifier and rule-based verifier.

Ablation: Effectiveness of Diverse Data

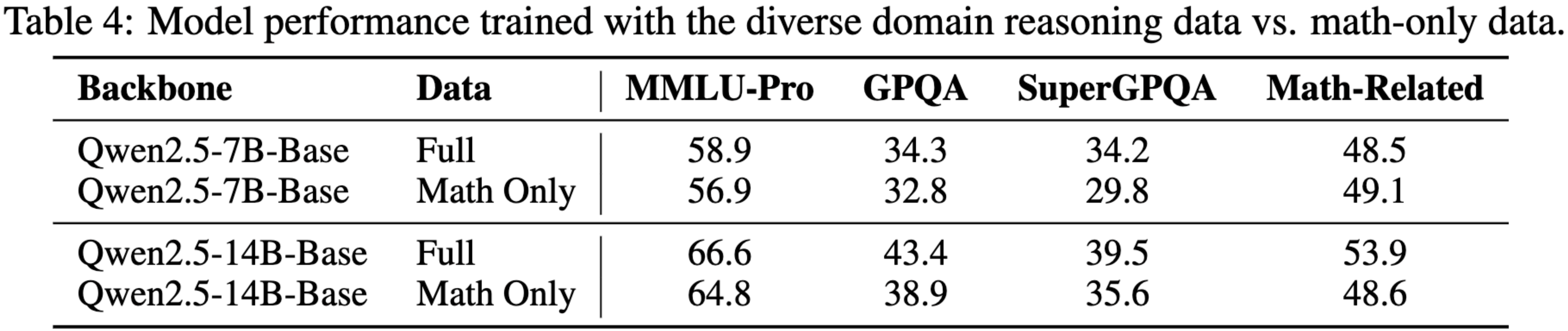


Table 4: Model performance trained with the diverse domain reasoning data vs. math-only data.

Backbone	**Data**	**MMLU-Pro**	**GPQA**	**SuperGPQA**	**Math-Related**
Qwen2.5-7B-Base	Full	58.9	34.3	34.2	48.5
Qwen2.5-7B-Base	Math Only	56.9	32.8	29.8	49.1
Qwen2.5-14B-Base	Full	66.6	43.4	39.5	53.9
Qwen2.5-14B-Base	Math Only	64.8	38.9	35.6	48.6

Takeaways: the Framework of Training

- Scaling up RLVR data is important to improve the reasoning capability of LLM.
- Model-based Verifier can make the scaling more effective.