General-Reasoner: Advancing LLM Reasoning Across All Domains

Xueguang Ma, Qian Liu, Dongfu Jiang, Wenhu Chen

Abstract

Existing open-sourced Deepseek-R1-Zero style reinforcement learning (RL) that directly trains base-LLM with reinforcement learning primarily focuses on mathematical reasoning that is easy to verify by a rule-based verifier. However, these constraints limit their applicability to broader domains with complex questions and long-tailed answer representations. In this work, we propose General-Reasoner, an approach aimed at enhancing LLM reasoning capabilities across diverse domains beyond mathematics. Our contributions include: (1) constructing a high-quality dataset of verifiable reasoning questions spanning a wide array of disciplines, and (2) developing a model-based answer verifier, which replaces traditional rule-based verification with a flexible, semantics-aware generative model. Our comprehensive evaluation across benchmarks such as MMLU-Pro, GPQA, and SuperGPQA demonstrates that General-Reasoner outperforms existing baseline methods, achieving robust and generalizable reasoning performance while maintaining superior effectiveness in mathematical reasoning tasks.

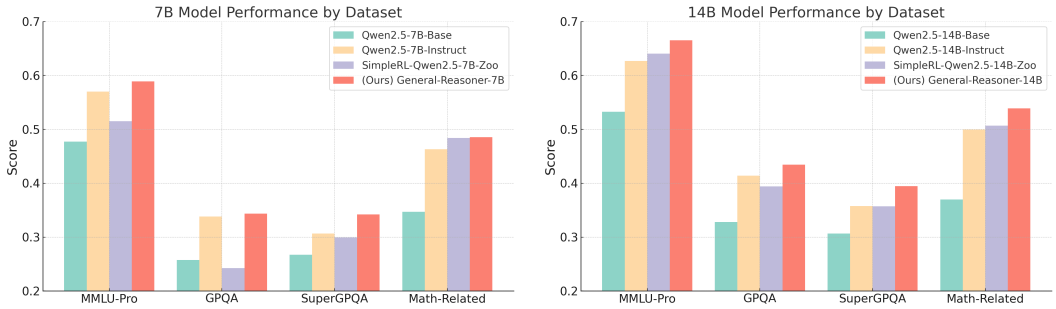


Figure 1: Effectiveness of our General-Reasoner compared to baseline methods on diverse language understanding and reasoning tasks.

1 Introduction

Recent advancements in reinforcement learning (RL) for large language models (LLMs) have demonstrated strong potential in enhancing reasoning capabilities. Notably, DeepSeek-R1-Zero [Team, 2025a] has shown that models trained purely through reinforcement learning – without intermediate supervised fine-tuning (SFT) – can achieve remarkable performance, inspiring a wave of open-source efforts replicating this "Zero" style training approach.

Several works, including SimpleRL [Zeng et al., 2025], DAPO [Yu et al., 2025], and DeepScaleR [Luo et al., 2025], have successfully adopted Group Relative Policy Optimization (GRPO) [Shao et al., 2024] or its variant to improve LLM reasoning. However, these efforts have largely focused on

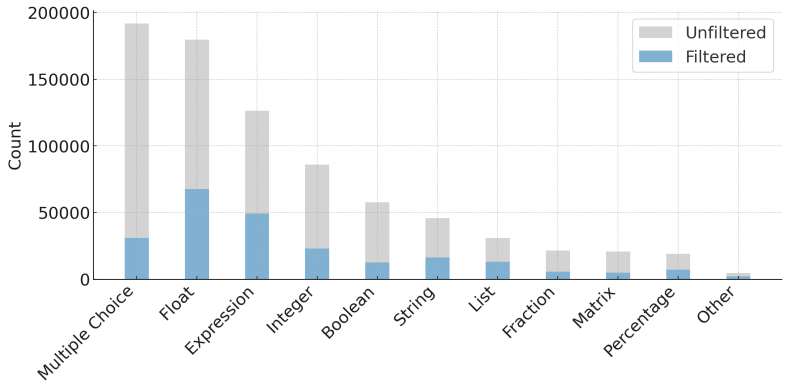


Figure 2: Distribution of answer types before and after filtering. The filtering process aims to reduce noise and trivial questions, ensuring diverse and challenging questions for improving models' reasoning capability.

mathematical reasoning tasks. This narrow scope limits the generalization of reasoning abilities to real-world scenarios where reasoning-heavy questions often span diverse domains and have answers not easy to capture by rule-based answer verification.

In this work, we propose a new paradigm for training LLMs to reason robustly and generalize across a wide range of domains beyond mathematics. We identify two key challenges in current approaches: (1) the lack of high-quality, publicly available, verifiable reasoning datasets from diverse fields, and (2) the restriction of rule-based verifiers, which struggle to provide accurate reward signals for varied answer expressions.

To address these challenges, we make two core contributions:

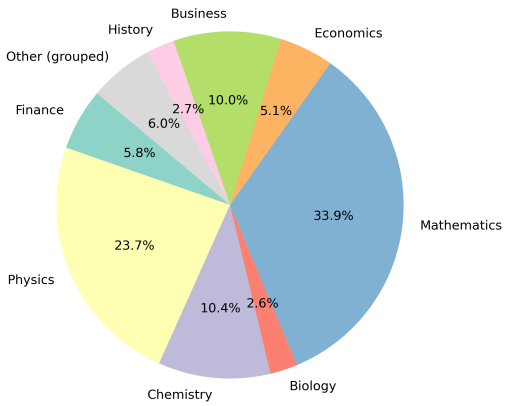
- **Diverse and Verifiable Dataset Collection:** We construct a large-scale dataset of high-quality reasoning questions across a wide range of domains, including science, economics, law, and the humanities. We use LLMs to select questions with verifiable answers, enabling reliable RL training across domains.
- **Model-Based Answer Verifier:** We introduce a compact 1.5B parameter generative model trained specifically for answer verification. This model replaces traditional rule-based systems, providing accurate and robust rewards across diverse answer types.

Together, these enable a scalable training paradigm for building LLMs with strong general reasoning capabilities. We validate our approach through comprehensive evaluations on benchmarks beyond math including MMLU-Pro [Wang et al., 2024], GPQA [Rein et al., 2023], and SuperGPQA [Team, 2025b]. Our model demonstrates consistent and significant performance gains over strong baselines.

2 Dataset

We construct a diverse, high-quality dataset to facilitate robust reasoning capabilities across a broad range of domains, extending beyond the commonly studied mathematical problems.

- We build upon the WebInstruct [Yue et al., 2024] dataset that covering broader domains of reasoning questions.
- Gemini-1.5-Pro [Team, 2024a] is utilized to selectively extract questions that have clearly verifiable answers, enhancing dataset reliability.
- Gemini-2.0-Flash then generates eight candidate answers per question to further filter the dataset:
  - We exclude questions unsolved by all eight Gemini-generated answers, thereby removing ambiguous or potentially noisy questions arising from web scraping.
  - We also exclude overly simple questions, where all eight Gemini-generated answers are correct, to maintain dataset complexity and ensure challenges for model generalization.



	Percentage
Other	3.15
Psychology	0.78
Computer Science	0.55
Health	0.5
Other STEM	0.42
Philosophy	0.22
Law	0.2
Engineering	0.14

Figure 3: Domain distribution of our curated 230K-question dataset, highlighting its balanced coverage across Mathematics, Physics, Chemistry, Finance, Economics, and other fields, enabling broader generalization in model reasoning capabilities. A breakdown of the grouped "Other" category shows its composition from various smaller disciplines.

	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

Table 1: Comparison of rule-based and model-based verifiers across different examples.

The distribution of answer types in our dataset is shown in Figure 2, highlighting a broad spectrum of answer formats such as multiple-choice, mathematical expressions. Through this careful filtering process, we curate a final dataset comprising approximately 230K questions that are challenging, verifiable, and diverse in nature. The distribution of the question domains of the curated dataset is shown in Figure 3.

3 Verifier

Rule-based verifiers are limited in scaling to broader domains and verifying more challenging questions due to the following reasons:

- **Rigid Matching Criteria:** Rule-based systems typically require an exact match or a predefined structure. This becomes problematic when semantically equivalent answers differ in expression.

Model Name	MMLU-Pro	GPQA	SuperGPQA
Qwen2.5-7B-Base	47.7	25.8	26.7
Qwen2.5-7B-Instruct	57.0	33.8	30.7
SimpleRL-Qwen2.5-7B-Zoo	51.5	24.2	29.9
General-Reasoner-7B	**58.9**	**34.3**	**34.2**
Qwen2.5-14B-Base	53.3	32.8	30.7
Qwen2.5-14B-Instruct	62.7	41.4	35.8
SimpleRL-Qwen2.5-14B-Zoo	64.0	39.4	35.7
General-Reasoner-14B	**66.6**	**43.4**	**39.5**

Table 2: Accuracy comparison across MMLU-Pro, GPQA, and SuperGPQA benchmarks.

Model Name	MATH-500	Olympiad	Minerva	GSM8K	AMC	AIME24x32	AIME25x32
Qwen2.5-7B-Base	60.2	28.6	36.0	83.1	30.0	3.8	1.4
Qwen2.5-7B-Instruct	75.0	39.4	45.2	90.9	52.5	12.5	8.5
SimpleRL-Qwen2.5-7B-Zoo	74.0	41.9	49.6	90.7	60.0	15.2	7.5
General-Reasoner-7B	76.0	37.9	54.0	92.7	55.0	13.8	10.4
Qwen2.5-14B-Base	65.4	33.5	24.3	91.6	37.5	3.6	2.9
Qwen2.5-14B-Instruct	77.4	44.7	52.2	94.5	57.5	12.2	11.0
SimpleRL-Qwen2.5-14B-Zoo	77.2	44.6	54.0	94.2	60.0	12.9	11.8
General-Reasoner-14B	78.6	42.1	58.1	94.2	70.0	17.5	16.9

Table 3: Accuracy on math-related tasks. For datasets other than AIME24 and AIME25, we utilize greedy decoding for a single round of generation. For AIME24 and AIME25, we calculate accuracy as an average across 32 runs using temperature$t = 1$, following the SimpleRL setting.

- **Inability to Interpret Semantics:** These systems cannot reason through different valid phrasings or units, especially in scientific or natural language answers.
- **Lack of Generality:** Rule-based systems can become a bottleneck in RL training for broader domains as they cannot easily adapt to the diverse representation norms in different disciplines.

To address these limitations, we introduce a **model-based verifier** – a compact 1.5B-parameter generative model fine-tuned specifically for answer verification. It is trained to assess whether a student-generated answer is equivalent to the ground truth in a generative manner, considering both the semantic context of the question and variability in valid expressions.

Specifically, we initialize the verifier using Qwen2.5-Math-1.5B [Yang et al., 2024]. The training data is constructed by prompting Gemini-2.0-Flash to verify the extracted short answers from its own generated solutions. The outputs, which include chain-of-thought justifications, serve as supervised fine-tuning targets for the 1.5B verifier. This approach enables the verifier to integrate seamlessly into our RL pipeline and reach approximately 90% agreement with Gemini-2.0-Flash.

4 Experiments

4.1 Training

We follow the "Zero" setting from recent works, directly conducting reinforcement learning (RL) from base large language models without an intermediate supervised fine-tuning stage.

Specifically, we initialize our models using Qwen2.5-7B and Qwen2.5-14B base models [Team, 2025c], and apply the GRPO algorithm.

Reward scores during training are calculated as follows:

- If the solution extraction fails (e.g., no boxed answer or summarization such as "the solution is:"), the reward is -0.5.
- If the solution passes verification, the base reward is 1, with a length-based penalty applied to discourage excessively long generations:

```
penalty = -0.05 × min(10, abs(length_of_ground_truth -
                    length_of_answer))
```

Training is conducted on 4 nodes with 8×H100 GPUs per node for approximately 700 steps. During training, the average model response length increases from approximately 700 tokens to around 1000 tokens. The total training time is around 2 days for the 7B model and around 4 days for the 14B model. Our implementation is based on the `verl` repository.[1]

4.2 Evaluation

To evaluate the models' general reasoning capabilities, we conduct a comprehensive assessment across several challenging benchmarks:

- **MMLU-Pro** [Wang et al., 2024]: A robust and challenging massive multi-task understanding dataset tailored to more rigorously benchmark large language models' capabilities.
- **SuperGPQA** [Team, 2025b]: A large-scale benchmark targeting graduate-level reasoning across 285 diverse disciplines.
- **GPQA** [Rein et al., 2023]: Graduate-level question answering designed to be resistant to shallow pattern-matching or memorization.
- **Math-Related Tasks**: A suite of standard math reasoning benchmarks, including MATH-500 [Hendrycks et al., 2021], Olympiad [He et al., 2024], Minerva [Lewkowycz et al., 2022], GSM8K [Cobbe et al., 2021], AMC, AIME24, and AIME25. We use the `simple-evals`[2] evaluation framework, and use GPT4o [Team, 2024b] to check the answer equivalence.

As shown in 2, our General-Reasoner models, both 7B and 14B, achieve strong performance across diverse domains. In particular, the improvements on MMLU-Pro and SuperGPQA highlight the model's robust reasoning capabilities in diverse and challenging settings, outperforming competitive baselines such as the SimpleRL-Zoo series and Qwen2.5-Instruct models.

Notably, this gain in general reasoning does not come at the cost of mathematical reasoning. In fact, General-Reasoner also shows superior performance on math-related tasks, demonstrating the effectiveness of our approach across domains as shown in Table 3.

Our method combines reinforcement learning with model-based verifiers, and diverse, high-quality reasoning data consistently improves performance over the base model by over 10 absolute points across the overall performance of all key metrics.

References

DeepSeek-AI Team. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025a.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Weinan Dai, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025.

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Li Erran Li, Raluca Ada Popa, and Ion Stoica. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl, 2025. Notion Blog.

[1] `https://github.com/volcengine/verl`

[2] `https://github.com/openai/simple-evals`

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.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

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 qa benchmark, 2023.

M-A-P Team. Supergpqa: Scaling llm evaluation across 285 graduate disciplines, 2025b.

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

Gemini Team. Gemini: A family of highly capable multimodal models, 2024a.

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, 2024. URL `https://arxiv.org/abs/2409.12122`.

Qwen Team. Qwen2.5 technical report, 2025c. URL `https://arxiv.org/abs/2412.15115`.

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*, 2021.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems, 2024.

Aitor Lewkowycz, Anders Johan Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Venkatesh Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

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

OpenAI Team. Gpt-4o system card, 2024b.

A Detailed Results

We provide detailed evaluation results on MMLU-pro, SuperGPQA are listed in below.

Model Name	**Avg**	**CS**	**Math**	**Chem**	**Eng**	**Law**	**Bio**	**Health**	**Phys**	**Bus**	**Phil**	**Econ**	**Other**	**Psy**	**Hist**
Qwen2.5-7B-Base	47.7	51.7	58.9	45.1	34.5	25.5	64.6	46.6	50.0	55.4	34.9	57.1	47.6	54.8	39.1
Qwen2.5-7B-Instruct	57.0	57.1	71.4	57.3	43.1	31.3	71.4	55.4	60.7	64.8	44.9	67.1	54.8	63.0	48.3
SimpleRL-Qwen2.5-7B-Zoo	51.5	54.9	55.4	48.7	40.9	30.8	68.3	53.1	53.4	58.0	41.3	62.6	52.2	60.7	42.5
General-Reasoner-7B	58.9	61.5	75.5	62.1	48.1	32.1	71.7	57.8	62.4	66.2	44.3	67.2	54.3	63.5	47.0
Qwen2.5-14B-Base	53.3	54.6	63.0	52.8	36.0	31.9	71.3	56.5	52.9	61.1	46.1	64.8	50.1	61.0	44.4
Qwen2.5-14B-Instruct	62.7	66.6	75.3	63.0	39.7	37.4	79.6	65.2	63.9	69.3	53.5	72.0	63.4	72.1	59.1
SimpleRL-Qwen2.5-14B-Zoo	64.0	66.1	75.8	66.9	49.8	37.2	79.5	64.1	67.7	69.5	55.5	73.8	61.4	70.9	53.8
General-Reasoner-Preview-14B	66.6	69.8	78.8	67.5	54.8	39.7	81.7	65.3	71.4	71.9	56.7	74.4	64.0	73.2	59.1

Table 4: Per-domain accuracy comparison of different models on MMLU-Pro.

Model Name	**Avg.**	Eng.	Med.	Sci.	Phil.	Mil. Sci.	Econ.	Mgmt.	Socio.	Lit./Arts	Hist.	Agron.	Law	Edu.
Qwen2.5-7B-Base	26.7	25.1	26.7	23.8	29.7	28.8	29.2	31.7	28.0	21.5	18.2	25.6	27.7	31.6
Qwen2.5-7B-Instruct	30.7	29.2	31.2	27.9	32.3	36.1	32.9	33.7	36.4	24.8	20.5	27.4	31.3	35.1
SimpleRL-Qwen2.5-7B-Zoo	29.9	28.0	31.3	26.0	34.9	32.2	32.7	31.5	34.3	25.0	23.1	27.2	29.4	33.5
General-Reasoner-7B	34.2	32.3	34.5	31.1	36.3	42.4	38.3	36.5	41.3	25.1	23.3	29.5	34.2	39.9
Qwen2.5-14B-Base	30.7	29.2	31.2	27.9	32.3	36.1	32.9	33.7	36.4	24.8	20.5	27.4	31.3	35.1
Qwen2.5-14B-Instruct	35.8	35.6	37.1	34.1	38.6	36.1	41.8	39.5	39.2	30.7	26.6	32.2	36.1	37.4
SimpleRL-Qwen2.5-14B-Zoo	35.7	34.2	36.7	33.0	36.0	40.0	39.9	41.3	38.5	30.8	26.7	31.1	38.1	38.2
General-Reasoner-Preview-14B	39.5	36.5	41.6	35.6	41.8	41.9	44.7	42.9	42.7	32.8	29.7	37.9	39.6	45.3

Table 5: Per-domain accuracy comparison of different models on SuperGPQA