DeepSeek-R1

DeepSeek-AI, Liang, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, Q., ... & Li, S. S. (2025). *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. arXiv preprint arXiv:2501.12948. https://doi.org/10.48550/arXiv.2501.12948

Nikita Ivanov &
Elias Abou Farhat

Table of Contents

Introduction

Reasoning

- The chain of thought or logical steps a model takes to reach a conclusion

Why do we care?

- It is key to decision making and aligns well with human preferences

DeepSeek-R1-Zero

Main goal:

- Want to explore potential of LLMs developing reasoning capabilities without any supervised data

Group Relative Policy Optimization (GRPO)

$$J_{\text{GRPO}}(\theta) = \left[\frac{1}{G} \sum_{i=1}^{G} \min \left(\overbrace{\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)} A_i}^{\text{The main thing we want to maximize}}, \overbrace{\text{clip}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1-\epsilon, 1+\epsilon \right) A_i}^{\text{If probability ratio is too far away from 1, clip it}} \right) - \overbrace{\beta D_{\text{KL}}\left(\pi_\theta \| \pi_{\text{ref}} \right)}^{\text{Regularization term}} \right]$$

$$A_i = \frac{r_i - \mathrm{mean}(\{r_1, r_2, \cdots, r_G\})}{\mathrm{std}(\{r_1, r_2, \cdots, r_G\})}.$$

DeepSeek-R1-Zero

Rewards & Data

- The reward is the source of the training signal
- Rule based reward based on:
  - **Accuracy**: Is the response correct?
  - **Format**: <think> Thinking Process </think>

- Stay away from neural reward model due to reward hacking

Data Template

A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> reasoning process here </think> <answer> answer here </answer>. User: prompt. Assistant:

DeepSeek-R1-Zero

Findings

- As DeepSeek evolves over time the model is able to generate more reasoning tokens which allows the model to explore and refine its thought process better
- The model is able to allocate more time to thinking and reevaluate its initial approach

Drawbacks

- Poor readability
- Language mixing

Question: If $a > 1$, then the sum of the real solutions of $\sqrt{a - \sqrt{a + x}} = x$ is equal to

Response: <think>
To solve the equation $\sqrt{a - \sqrt{a + x}} = x$, let's start by squaring both $\cdots$
$\left(\sqrt{a - \sqrt{a + x}}\right)^2 = x^2 \implies a - \sqrt{a + x} = x^2.$
Rearrange to isolate the inner square root term:
$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$
$\ldots$
Wait, wait. Wait. That's an aha moment I can flag here.
Let's reevaluate this step-by-step to identify if the correct sum can be $\cdots$
We started with the equation:
$\sqrt{a - \sqrt{a + x}} = x$
First, let's square both sides:
$a - \sqrt{a + x} = x^2 \implies \sqrt{a + x} = a - x^2$
Next, I could square both sides again, treating the equation: $\cdots$
$\ldots$

DeepSeek-R1

Main questions:

- *Can reasoning performance be further improved or convergence accelerated by incorporating a small amount of high-quality data as a cold start?*
- *How can we train a user-friendly model that not only produces clear and coherent Chain of Thoughts but also demonstrates strong general capabilities?*

Pipeline:

- Cold Start
- Reasoning Oriented RL
- Rejection Sampling and SFT
- RL for All Scenarios

DeepSeek-R1 Cold Start

Cold Start

- Aimed to avoid early unstable cold start phase of RL training from base model
- Collected small amount of long CoT data to finetune the model initially
- To collect the data
    - Using few shot prompting with long CoT as an example
    - Directly prompting models to generate detailed answers
    - Gathering DeepSeek-R1-Zero outputs in a readable format and using human annotators

Advantages of Cold Start
- **Readability**: DeepSeek-R1-Zero content is often not suitable for reading
- **Potential**: Careful, human designed data performs better against DeepSeek-R1-Zero

DeepSeek-R1

Reasoning Oriented RL

- Focus on enhancing models reasoning capabilities for
  - Coding
  - Math
  - Science
  - Logic and reasoning

Language Mixing

- Introduce language consistency reward during RL training
- Proportion of target language words in the chain of thought
- This actually degrades the models performance but aligns with human preference making it more readable

DeepSeek-R1

Rejection Sampling and SFT

- This stage incorporates data from other domains to enhance model writing, role-playing and other tasks

Reasoning Data

- Created reasoning prompts and reasoning trajectories by rejection sampling from checkpoint
- Incorporate additional data which may use generative reward model
    - Feed ground-truth and model predictions into DeepSeek-V3 for judgement

Non-Reasoning Data

- Reuse portions of the SFT dataset for DeepSeek-V3
- Call DeepSeek-V3 to generate potential CoT before answering query by prompting

DeepSeek-R1

Rl for All Scenarios

- Aim is to align model with human preferences, helpfulness and harmlessness
- Train model using combination of reward signals and diverse prompts
  - For reasoning data use rule based reward
  - For general data use reward model to capture human preferences
- For helpfulness
  - Focus on the final summary
- For harmlessness
  - Evaluate entire response to identify any potential risks, biases or harmful content generated

Distillation

Main goal:

- Equip smaller models with reasoning capabilities like DeepSeek-R1

Distillation vs RL

- Shown that bigger model distilled into smaller is much better than smaller model with large scaled RL

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCodeBench
	pass@1	cons@64	pass@1	pass@1	pass@1
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9
DeepSeek-R1-Zero-Qwen-32B	47.0	60.0	91.6	55.0	40.2
DeepSeek-R1-Distill-Qwen-32B	**72.6**	**83.3**	**94.3**	**62.1**	**57.2**

Experiments/Results

DeepSeek-R1-Zero Results

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	1820
OpenAI-o1-0912	74.4	83.3	94.8	77.3	63.4	1843
DeepSeek-R1-Zero	71.0	86.7	95.9	73.3	50.0	1444

Distillation Results

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	**1820**
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9	1316
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481
DeepSeek-R1-Distill-Qwen-32B	**72.6**	83.3	94.3	62.1	57.2	1691
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205
DeepSeek-R1-Distill-Llama-70B	70.0	**86.7**	**94.5**	**65.2**	**57.5**	1633

DeepSeek-R1 Results

	Benchmark (Metric)	Claude-3.5-Sonnet-1022	GPT-4o 0513	DeepSeek V3	OpenAI o1-mini	OpenAI o1-1217	DeepSeek R1
	Architecture	-	-	MoE	-	-	MoE
	# Activated Params	-	-	37B	-	-	37B
	# Total Params	-	-	671B	-	-	671B
English	MMLU (Pass@1)	88.3	87.2	88.5	85.2	**91.8**	90.8
	MMLU-Redux (EM)	88.9	88.0	89.1	86.7	-	**92.9**
	MMLU-Pro (EM)	78.0	72.6	75.9	80.3	-	**84.0**
	DROP (3-shot F1)	88.3	83.7	91.6	83.9	90.2	**92.2**
	IF-Eval (Prompt Strict)	**86.5**	84.3	86.1	84.8	-	83.3
	GPQA Diamond (Pass@1)	65.0	49.9	59.1	60.0	**75.7**	71.5
	SimpleQA (Correct)	28.4	38.2	24.9	7.0	**47.0**	30.1
	FRAMES (Acc.)	72.5	80.5	73.3	76.9	-	**82.5**
	AlpacaEval2.0 (LC-winrate)	52.0	51.1	70.0	57.8	-	**87.6**
	ArenaHard (GPT-4-1106)	85.2	80.4	85.5	92.0	-	**92.3**
Code	LiveCodeBench (Pass@1-COT)	38.9	32.9	36.2	53.8	63.4	**65.9**
	Codeforces (Percentile)	20.3	23.6	58.7	93.4	**96.6**	96.3
	Codeforces (Rating)	717	759	1134	1820	**2061**	2029
	SWE Verified (Resolved)	**50.8**	38.8	42.0	41.6	48.9	49.2
	Aider-Polyglot (Acc.)	45.3	16.0	49.6	32.9	**61.7**	53.3
Math	AIME 2024 (Pass@1)	16.0	9.3	39.2	63.6	79.2	**79.8**
	MATH-500 (Pass@1)	78.3	74.6	90.2	90.0	96.4	**97.3**
	CNMO 2024 (Pass@1)	13.1	10.8	43.2	67.6	-	**78.8**
Chinese	CLUEWSC (EM)	85.4	87.9	90.9	89.9	-	**92.8**
	C-Eval (EM)	76.7	76.0	86.5	68.9	-	**91.8**
	C-SimpleQA (Correct)	55.4	58.7	**68.0**	40.3	-	63.7

Notebook