



DEEPSEEK-R1

Exploring Cutting-Edge Technology Shaping The Future

Ekaterina Lipina

1 April, 2025

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

DeepSeek-AI

research@deepseek.com

Abstract

We introduce our first-generation reasoning models, DeepSeek-R1-Zero and DeepSeek-R1. DeepSeek-R1-Zero, a model trained via large-scale reinforcement learning (RL) without supervised fine-tuning (SFT) as a preliminary step, demonstrates remarkable reasoning capabilities. Through RL, DeepSeek-R1-Zero naturally emerges with numerous powerful and intriguing reasoning behaviors. However, it encounters challenges such as poor readability, and language mixing. To address these issues and further enhance reasoning performance, we introduce DeepSeek-R1, which incorporates multi-stage training and cold-start data before RL. DeepSeek-R1 achieves performance comparable to OpenAI-o1-1217 on reasoning tasks. To support the research community, we open-source DeepSeek-R1-Zero, DeepSeek-R1, and six dense models (1.5B, 7B, 8B, 14B, 32B, 70B) distilled from DeepSeek-R1 based on Qwen and Llama.

WHAT YOU'LL LEARN TODAY



1.

WHAT MAKES DEEPSEEK-R1 DIFFERENT FROM OTHER LLMS?

2.

WHY THE MOVE TO PURE RL TRAINING MADE IT SO REVOLUTIONARY?

3.

HOW THE MODEL WAS TRAINED TO REASON?

AGENDA

1. LLM Training process
2. What is Reinforcement Learning?
3. Chain-of-Thought Prompting
4. Reasoning Models and OpenAI's o1
5. DeepSeek-R1 Architecture
6. R1 vs. R1-Zero
7. Group Relative Policy Optimization (GRPO)
8. Aha Moment
9. Drawback of R1-Zero
10. R1 Training Procedure
11. Results

LLM TRAINING PROCESS

PRE-TRAINING

Next token prediction

Write a bedtime _____

↑
story

SUPERVISED FINE-TUNING

Fine-tune

Instructions Dataset

(instruction, response)

SUPERVISED FINE-TUNING

Reinforcement Learning from Human Feedback (RLHF)

Reinforcement Learning from AI Feedback

LLM TRAINING PROCESS

PRE-TRAINING

Next token prediction

Write a bedtime _____

↑
story

~~SUPERVISED FINE-TUNING~~

~~Fine-tune~~

~~Instructions Dataset~~

~~(instruction, response)~~

~~SUPERVISED FINE-TUNING~~

~~Reinforcement Learning from Human Feedback (RLHF)~~

~~Reinforcement Learning from AI Feedback~~

Rule-based Reinforcement Learning

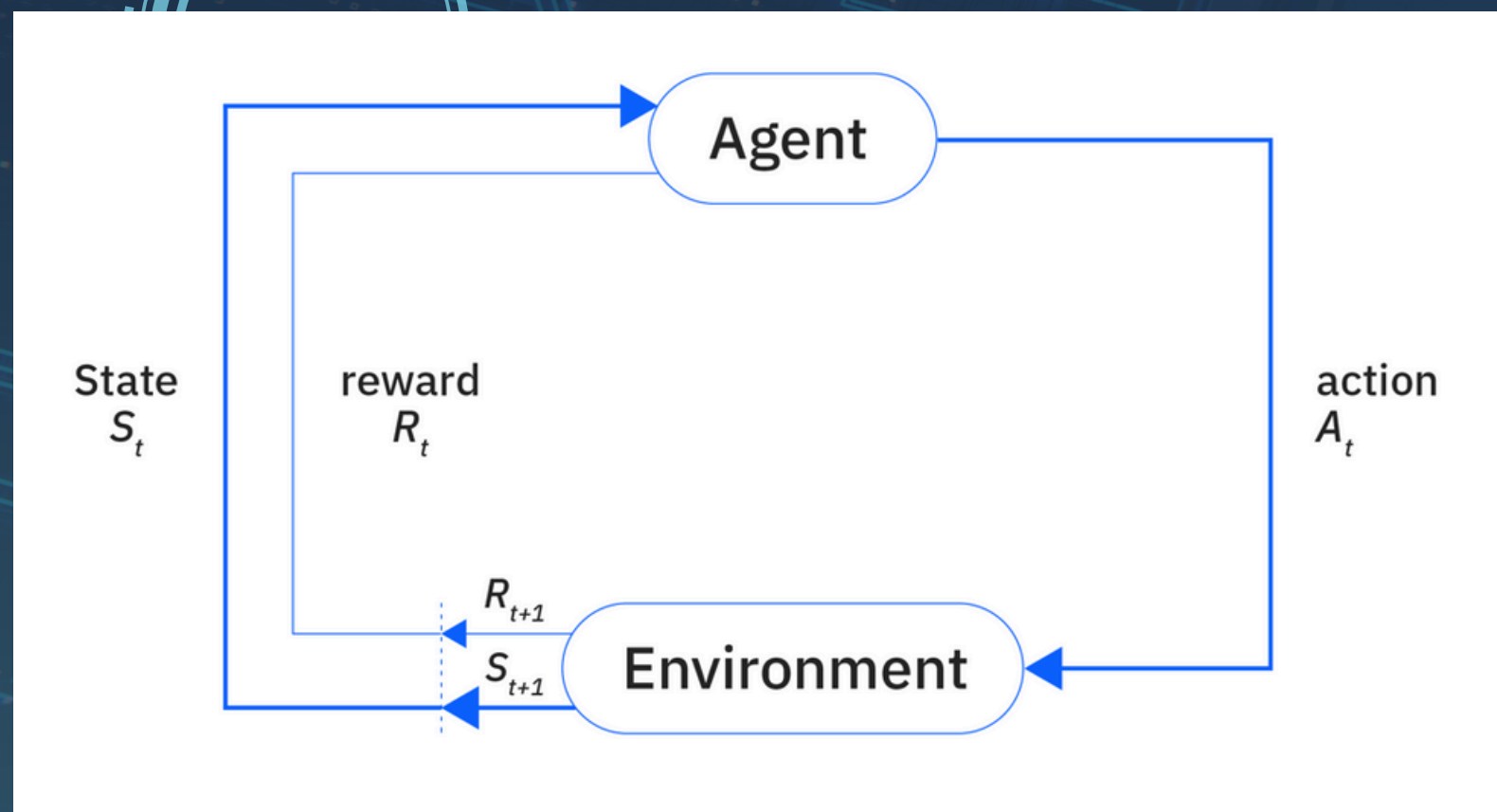
WHAT IS REINFORCEMENT LEARNING?

Agent: Learns and makes decisions

Environment: Where the agent operates

The agent sees the **current state S_t** , chooses an **action A_t** , the environment responds with a **new state S_{t+1}** and a **reward R_t** , and the process repeats.

Agent learns a **policy** — a strategy that helps it pick the best actions to maximize cumulative reward



CHAIN-OF-THOUGHT (COT)

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅

OPENAI HIDDEN COT FOR O1 MODEL

Hiding the Chains of Thought

We believe that a hidden chain of thought presents a unique opportunity for monitoring models. Assuming it is faithful and legible, the hidden chain of thought allows us to "read the mind" of the model and understand its thought process. For example, in the future we may wish to monitor the chain of thought for signs of manipulating the user. However, for this to work the model must have freedom to express its thoughts in unaltered form, so we cannot train any policy compliance or user preferences onto the chain of thought. We also do not want to make an unaligned chain of thought directly visible to users.

Therefore, after weighing multiple factors including user experience, competitive advantage, and the option to pursue the chain of thought monitoring, we have decided not to show the raw chains of thought to users. We acknowledge this decision has disadvantages. We strive to partially make up for it by teaching the model to reproduce any useful ideas from the chain of thought in the answer. For the o1 model series we show a model-generated summary of the chain of thought.

REASONING MODELS AND OPENAI'S O1

Earlier models like GPT-4

think for **too long** →

accuracy **decreases**

hallucinations

"lost in the middle" problem
– the model forgets the center
of the prompt

Latest generation of models
(like o1, R1, and others)

think for **longer** →

accuracy **increases**

"test-time scaling" phenomenon
– improve performance just by
making it reason more deeply at
inference time (instead of
pretraining)

DEEPSEEK-R1 OVERVIEW

Reasoning language model

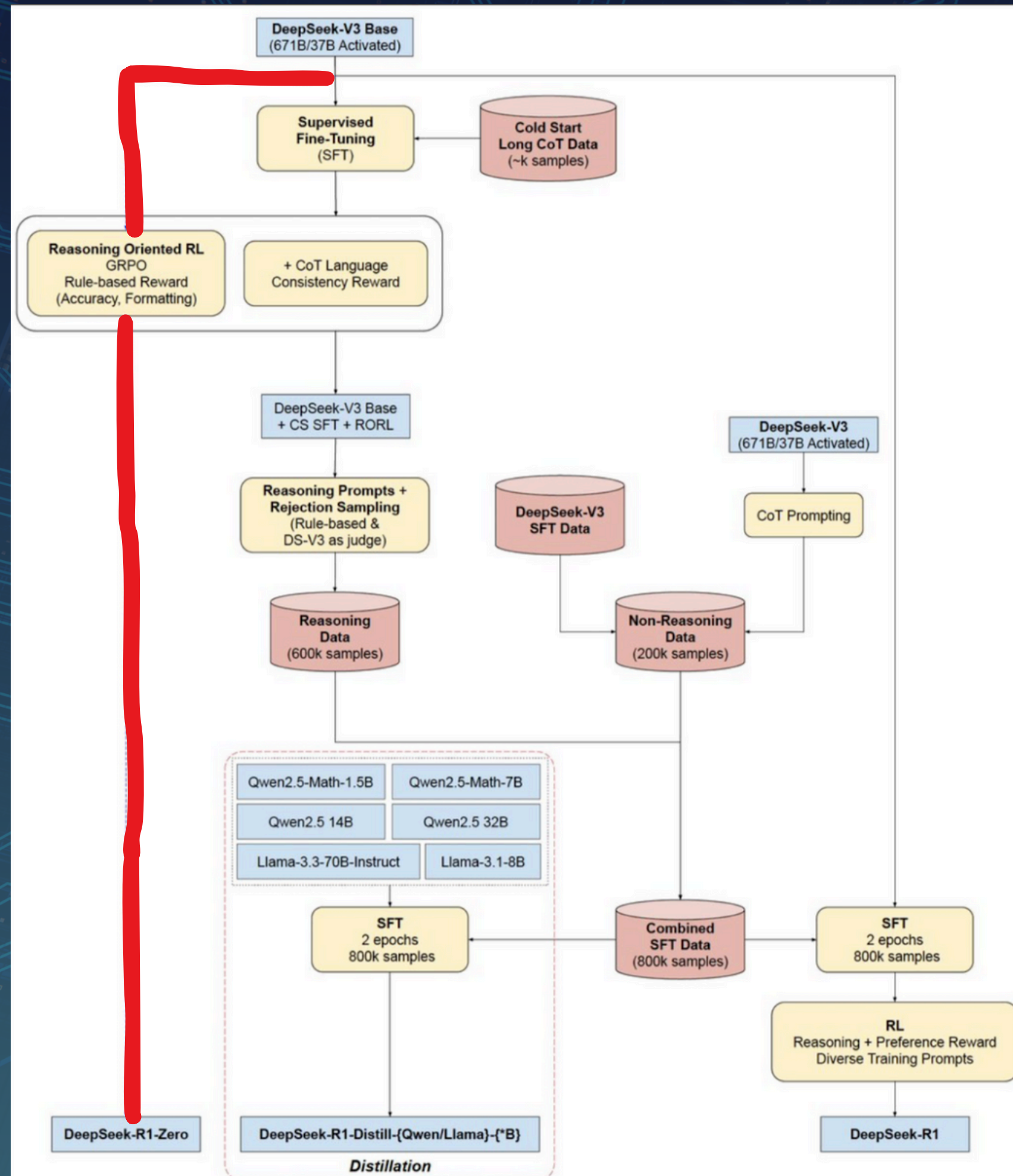
Builds on top of the **DeepSeek-V3** architecture

Mixture-of-Experts (MoE) - activate only a subset of parameters during each forward pass

671 billion parameters total

~37 billion active parameters per token

Transformer decoder architecture



TRAINING: R1 VS R1-ZERO

R1-Zero

no SFT (Supervised Fine-Tuning)

no human-labeled examples involved

"pure reinforcement learning" =
no human in the loop (GRPO)

rewards

unit tests (for programming tasks)
mathematical validations
constraints

= rules

R1

multi-stage

traditional setup

SFT+RL



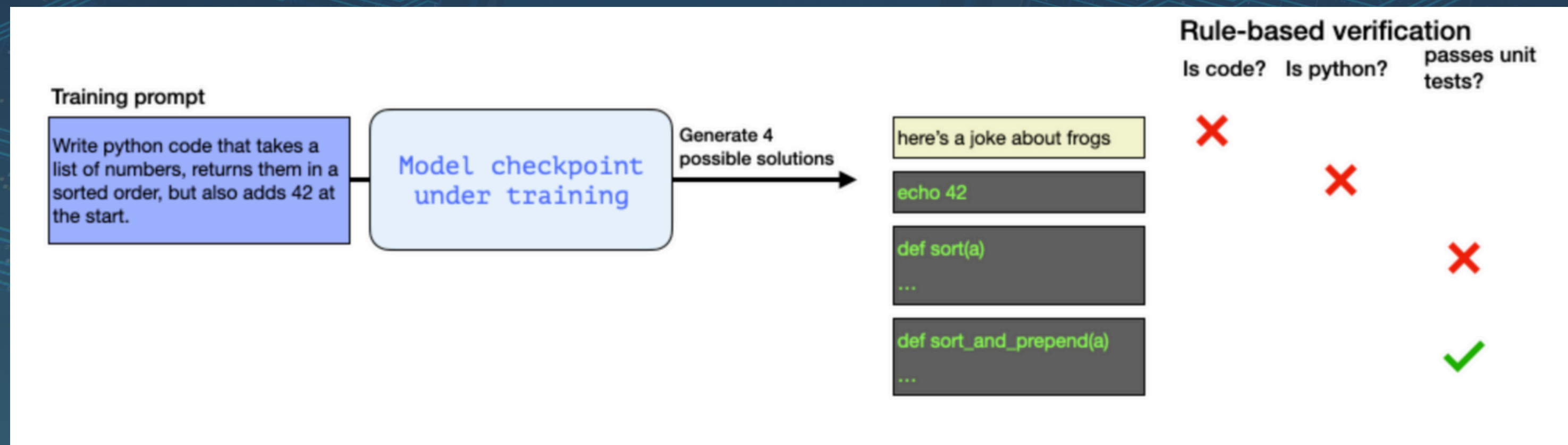
GROUP RELATIVE POLICY OPTIMIZATION (GRPO)

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^G \sim \pi_{\theta_{old}}(O|q)]$$
$$\frac{1}{G} \sum_{i=1}^G \left(\min \left(\frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{old}}(o_i|q)} A_i, \text{clip} \left(\frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1 - \varepsilon, 1 + \varepsilon \right) A_i \right) - \beta \mathbb{D}_{KL}(\pi_{\theta} \parallel \pi_{ref}) \right), \quad (1)$$

$$\mathbb{D}_{KL}(\pi_{\theta} \parallel \pi_{ref}) = \frac{\pi_{ref}(o_i|q)}{\pi_{\theta}(o_i|q)} - \log \frac{\pi_{ref}(o_i|q)}{\pi_{\theta}(o_i|q)} - 1, \quad (2)$$

where ε and β are hyper-parameters, and A_i is the advantage, computed using a group of rewards $\{r_1, r_2, \dots, r_G\}$ corresponding to the outputs within each group:

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \dots, r_G\})}{\text{std}(\{r_1, r_2, \dots, r_G\})}. \quad (3)$$



AHA MOMENT

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 ...

$$(\sqrt{a - \sqrt{a+x}})^2 = x^2 \implies a - \sqrt{a+x} = x^2.$$

Rearrange to isolate the inner square root term:

$$(a - x^2)^2 = a + x \implies a^2 - 2ax^2 + (x^2)^2 = a + x \implies x^4 - 2ax^2 - x + (a^2 - a) = 0$$

...

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 ...

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: ...

...

Table 3 | An interesting "aha moment" of an intermediate version of DeepSeek-R1-Zero. The model learns to rethink using an anthropomorphic tone. This is also an aha moment for us, allowing us to witness the power and beauty of reinforcement learning.



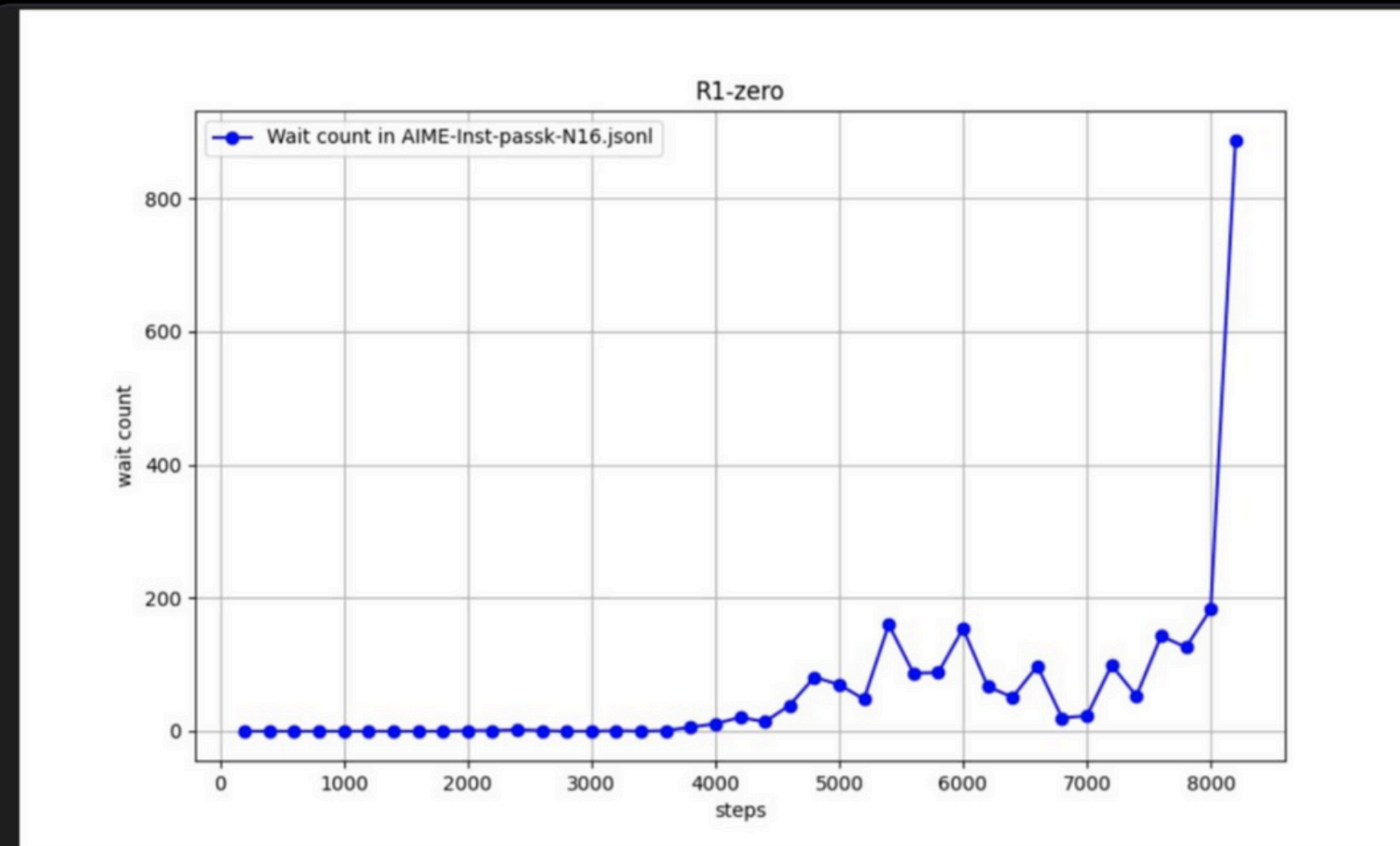
Peiyi Wang

@sybilhyz



During RL training, the model's reasoning patterns evolve continuously. At times, a specific pattern may suddenly emerge prominently, which I define as the "aha moment". For instance, the image below illustrates the emergency of the `wait` pattern in one of my experiments.

[Перевести пост](#)



8:22 PM · 28 янв. 2025 г. · **18,7 тыс.** просмотра

RESPONSE LENGTH

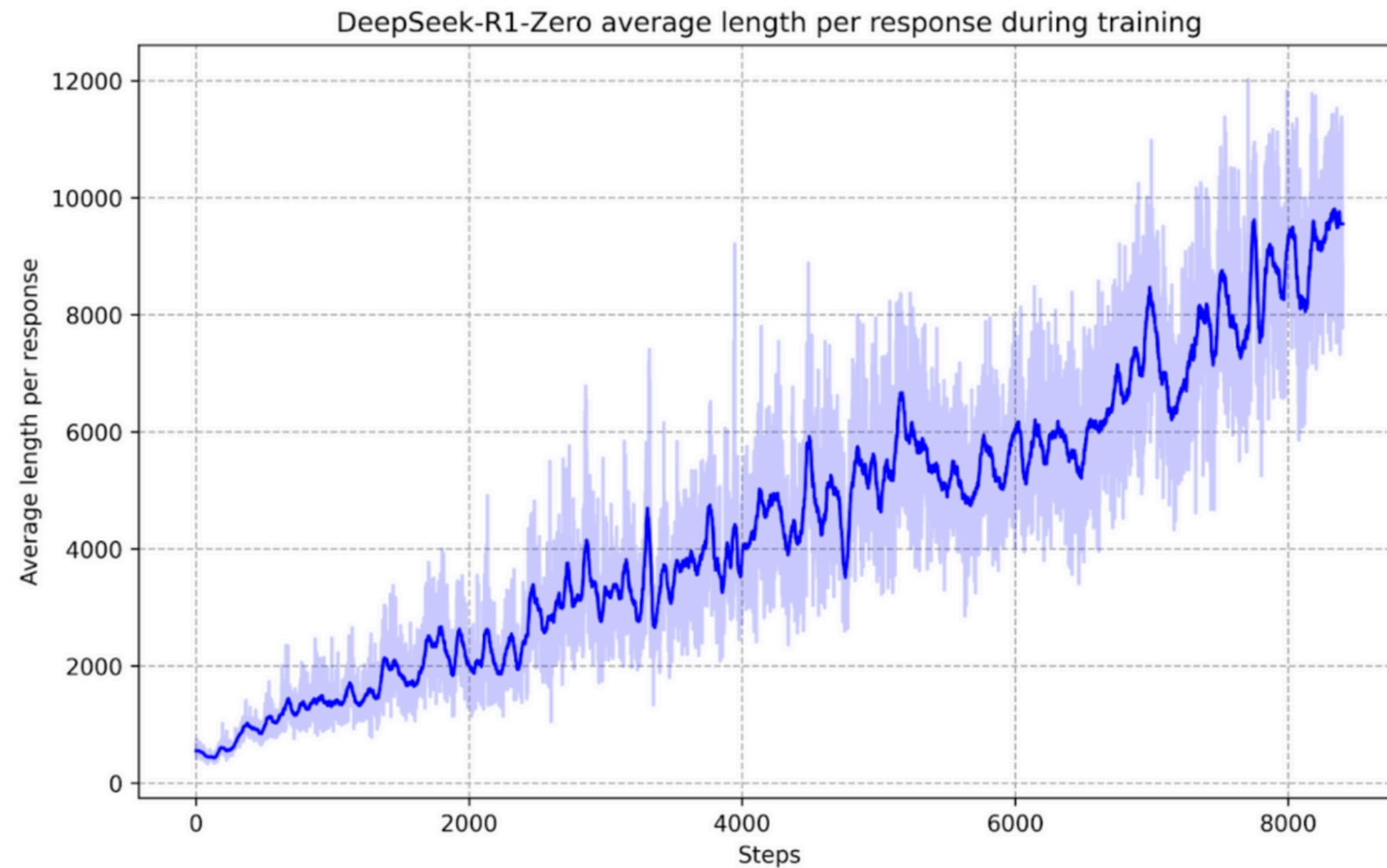


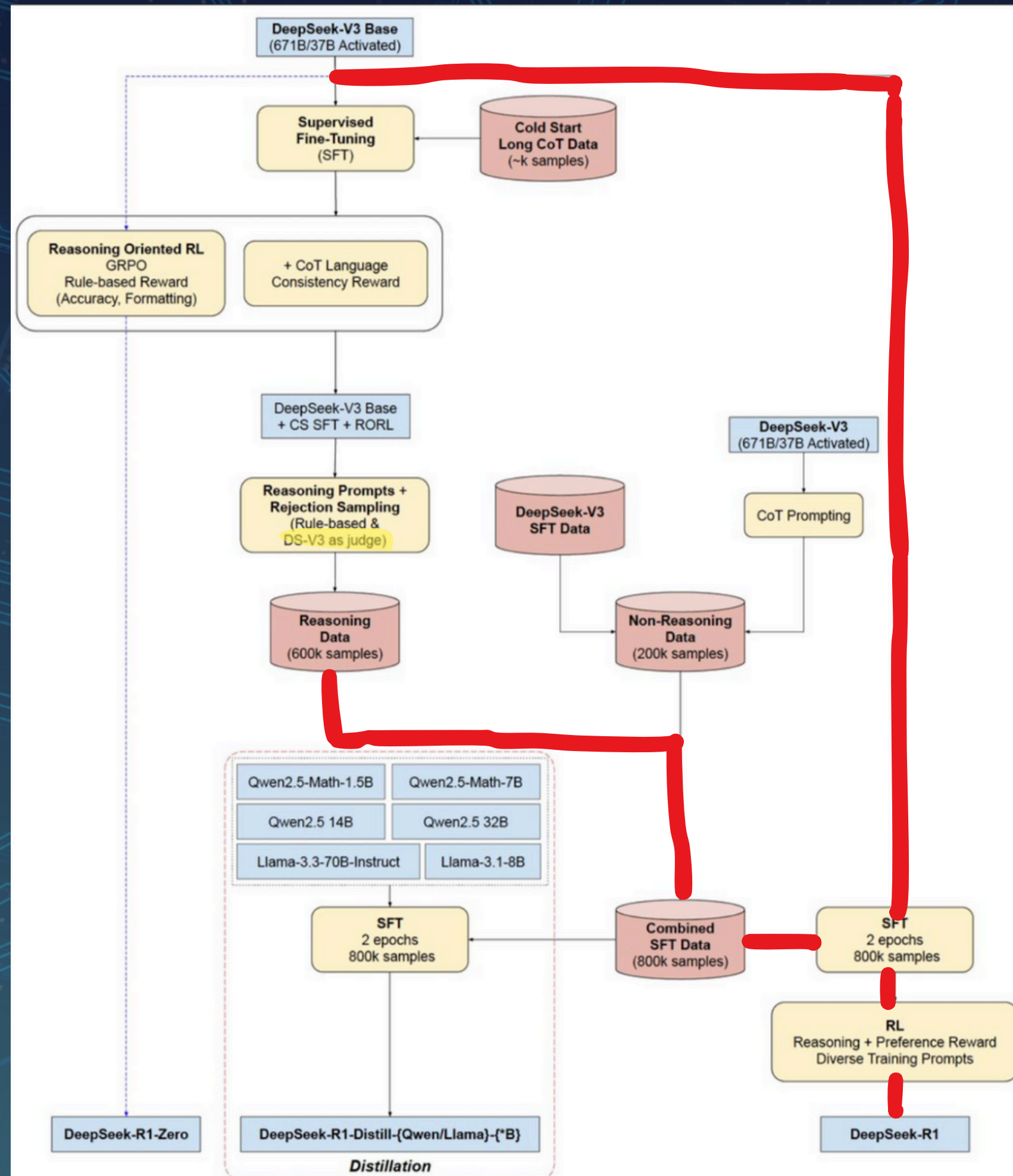
Figure 3 | The average response length of DeepSeek-R1-Zero on the training set during the RL process. DeepSeek-R1-Zero naturally learns to solve reasoning tasks with more thinking time.

DRAWBACK OF R1-ZERO

Drawback of DeepSeek-R1-Zero Although DeepSeek-R1-Zero exhibits strong reasoning capabilities and autonomously develops unexpected and powerful reasoning behaviors, it faces several issues. For instance, DeepSeek-R1-Zero struggles with challenges like poor readability, and language mixing. To make reasoning processes more readable and share them with the open community, we explore DeepSeek-R1, a method that utilizes RL with human-friendly cold-start data.

poor readability
language mixing

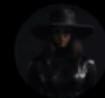
back to SFT → R1 model



HELPFULNESS & HARMLESSNESS

2.3.4. Reinforcement Learning for all Scenarios

To further align the model with human preferences, we implement a secondary reinforcement learning stage aimed at improving the model's helpfulness and harmlessness while simultaneously refining its reasoning capabilities. Specifically, we train the model using a combination of reward signals and diverse prompt distributions. For reasoning data, we adhere to the methodology outlined in DeepSeek-R1-Zero, which utilizes rule-based rewards to guide the learning process in math, code, and logical reasoning domains. For general data, we resort to reward models to capture human preferences in complex and nuanced scenarios. We build upon the DeepSeek-V3 pipeline and adopt a similar distribution of preference pairs and training prompts. For helpfulness, we focus exclusively on the final summary, ensuring that the assessment emphasizes the utility and relevance of the response to the user while minimizing interference with the underlying reasoning process. For harmlessness, we evaluate the entire response of the model, including both the reasoning process and the summary, to identify and mitigate any potential risks, biases, or harmful content that may arise during the generation process. Ultimately, the integration of reward signals and diverse data distributions enables us to train a model that excels in reasoning while prioritizing helpfulness and harmlessness.



Flowers 
@flowersslop



DeepSeek R1 when it is prompted to select a random number:

Okay, the user wants me to think of a number between 1 and 10,000 but not tell them. Hmm, they mentioned not to overthink it, so maybe just pick a random number quickly. Let me see. Well, people often have favorite numbers or patterns. Maybe go with something like 7777? But that's four 7s, which might be too obvious. Alternatively, 1234? Too sequential. Maybe a number that's a common choice but not too predictable. Wait, I should just pick the first number that comes to mind. Let me try that. Closing my eyes and... 42? Wait, that's from Hitchhiker's Guide, but maybe they don't want a reference. Alternatively, 1004? That's within the range. Or 8675? No, that's from a song. Hmm, maybe 3141? Like the first four digits of pi. But 3141 is 3.141... multiplied by 1000. That's clever, but maybe too nerdy. Alternatively, 1984? But that's a book title. Maybe 2023, the current year? But they might not want a recent year. Or 1337, which is leet speak. But that's only 4 digits. Wait, 1337 is 1-3-3-7, which is 1337. Maybe that's too slangy. Alternatively, 5000, the midpoint. But that's too obvious. How about 777? No, that's three digits. Wait, the number needs to be between 1 and 10,000, so up to four digits. Let me just pick 42 again. Wait, that's too low. Maybe 420? But that's associated with cannabis culture. Not sure. Maybe 6969? That's a playful number but might be seen as immature. Alternatively, 1111? A repeating digit. Or 2222? Hmm. Maybe 1234 is too simple. Wait, the user said not to overthink, so just pick a number without second-guessing. Let me go with 7321. No reason, just random digits. But maybe that's too arbitrary. Alternatively, 500, but 0500 is 500. Wait, leading zeros don't count. So 500 is 500. Maybe 8888? That's considered lucky in some cultures. But again, maybe too obvious. Alright, I'm overcomplicating it. The user said not to overthink. Let me just pick the first number that comes to mind. Okay, 7777. Done.

[Перевести пост](#)

4:40 AM · 23 янв. 2025 г. · **498,1 тыс.** просмотра

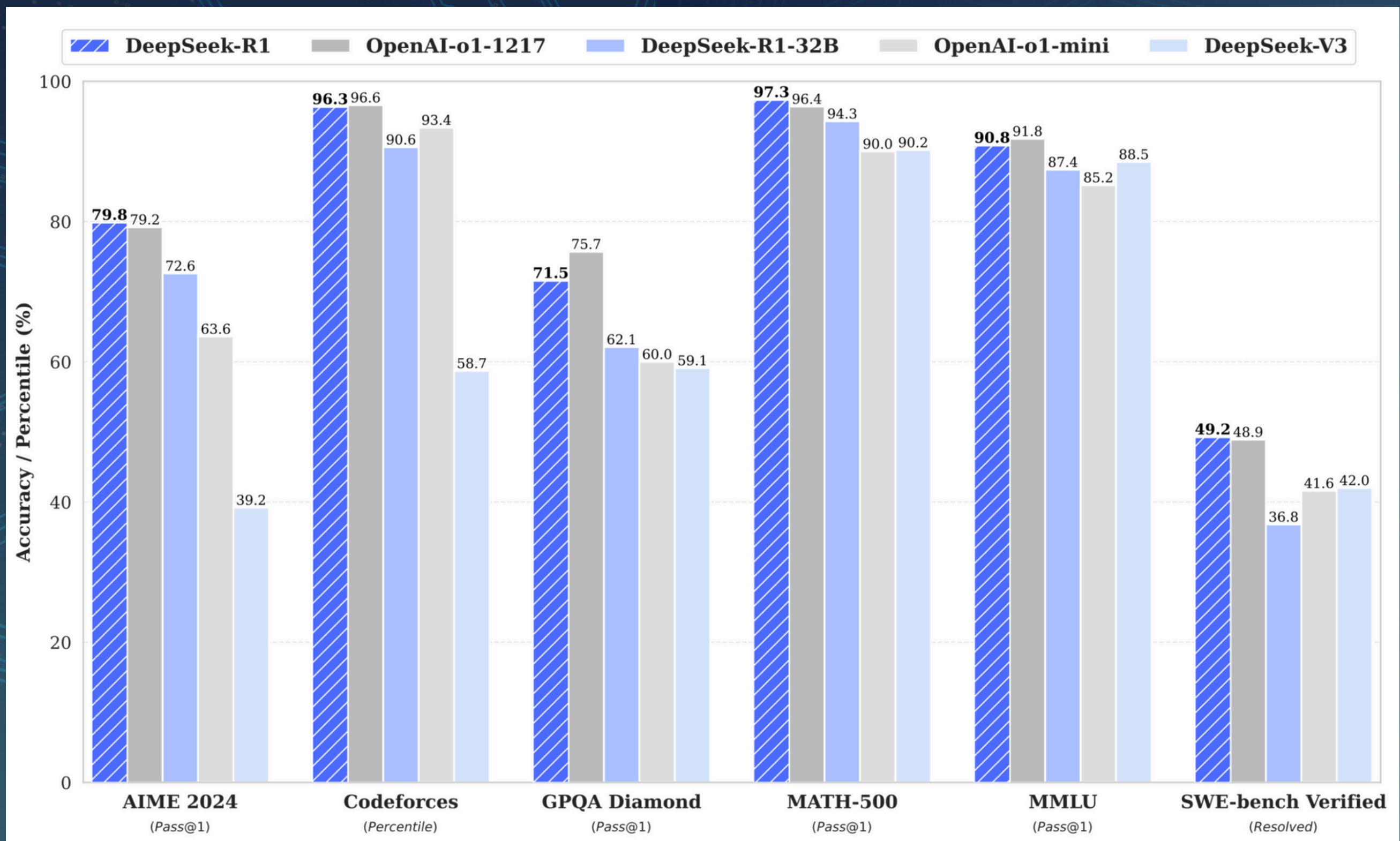
RESULTS

3.1. DeepSeek-R1 Evaluation

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

Table 4 | Comparison between DeepSeek-R1 and other representative models.

COMPARISION





THANK YOU!
QUESTIONS?