

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

Author: DeepSeek-AI

Presented by Kunzhao Xu and Yifei Liu
Thanks to Chaoyi Ruan's Guidance



Agenda

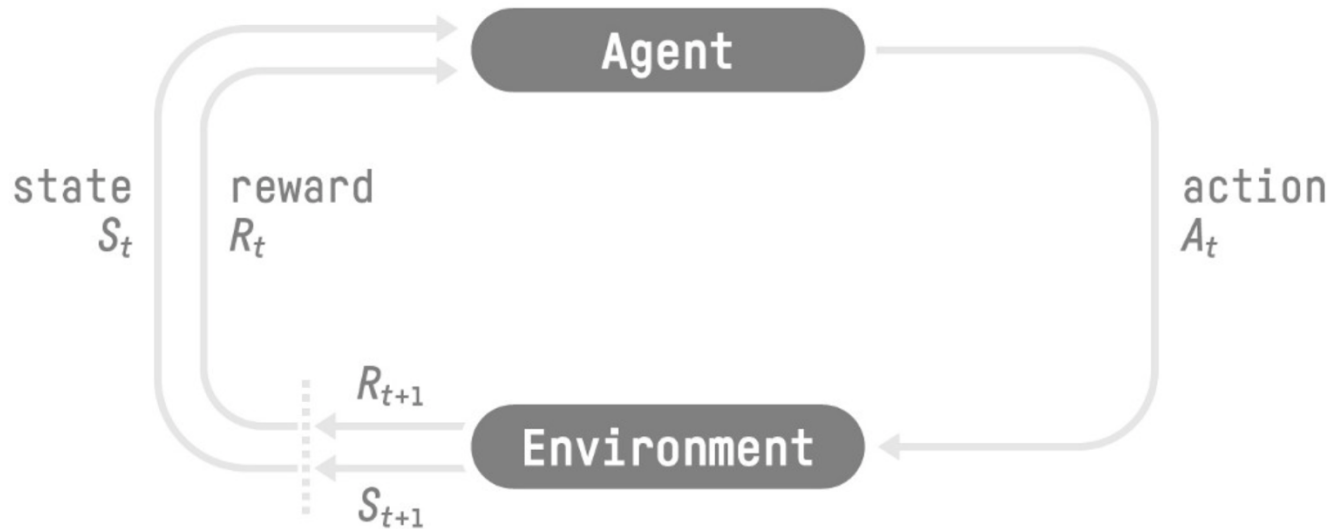
- Post-training
- Background information about RL
- DeepSeek-R1-Zero
- DeepSeek-R1

- Why post-training matters
 - Task specialization
 - Test-time scaling
 - Alignment
- Methods
 - Process-based reward model (PRM)
 - Reinforcement learning
 - Searching (MCTS, beam search etc.)

- Why post-training matters
 - Task specialization
 - Test-time scaling
 - Alignment
- Methods
 - Process-based reward model (PRM)
 - Reinforcement learning
 - Searching (MCTS, beam search etc.)

Background information about RL

- What is RL



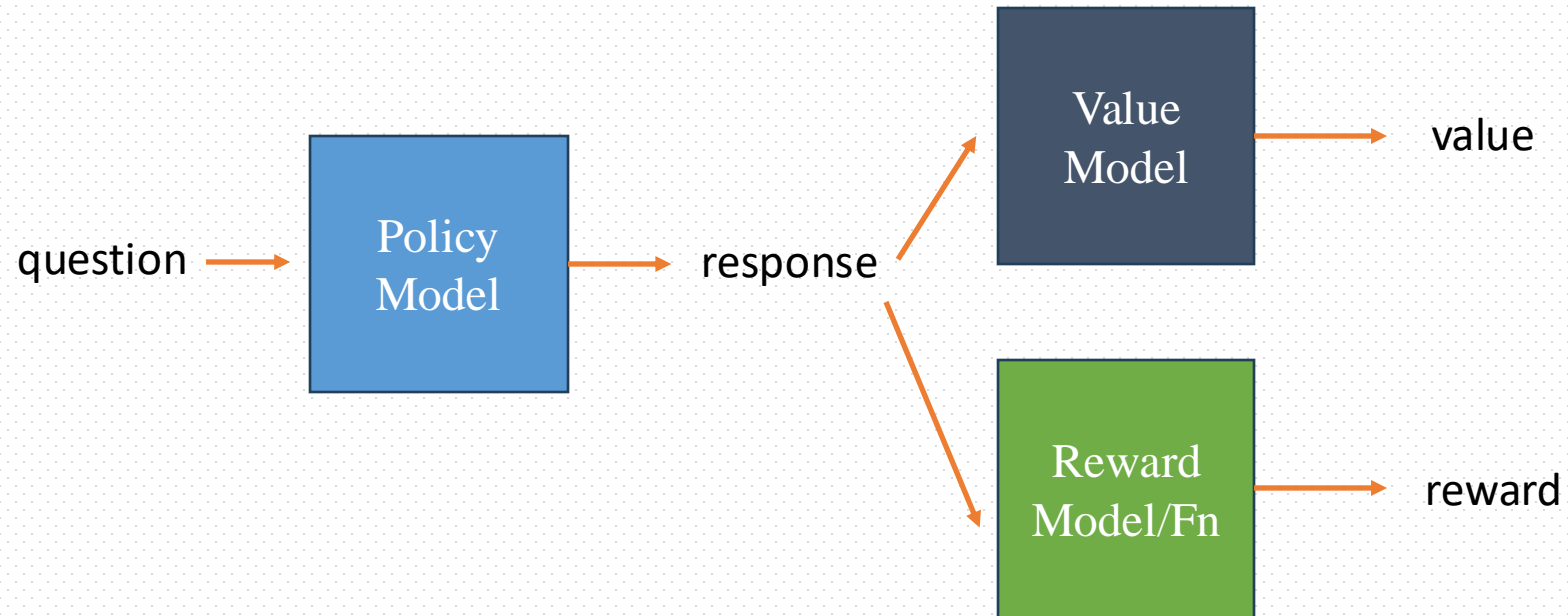
- Goal: maximize cumulative reward
- Snake game as an example

Background information about RL



- Policy (π): decide which action to take given the state we are in
- Reward (r_t): a scalar feedback signal that quantifies the immediate desirability of an agent's action in a specific state
 - Two type
 - Rule-based
 - Model-based
 - Two granularity
 - Outcome-Reward Model (ORM)
 - Process-Reward Model (PRM)
- Value ($V^\pi(s_t)$): estimate the expected cumulative reward

Background information about RL



- Proximal Policy Optimization (PPO)
 - Goal: improves our agent's training stability by avoiding policy updates that are too large
 - Objective

$$\mathcal{J}_{PPO}(\theta) = \mathbb{E}[q \sim P(Q), o \sim \pi_{\theta_{old}}(O|q)] \frac{1}{|o|} \sum_{t=1}^{|o|} \min \left[\frac{\pi_{\theta}(o_t|q, o_{<t})}{\pi_{\theta_{old}}(o_t|q, o_{<t})} A_t, \text{clip} \left(\frac{\pi_{\theta}(o_t|q, o_{<t})}{\pi_{\theta_{old}}(o_t|q, o_{<t})}, 1 - \epsilon, 1 + \epsilon \right) A_t \right]$$

Background information about RL

- Proximal Policy Optimization (PPO)
 - Goal: improves our agent's training stability by avoiding policy updates that are too large
 - Objective

$$\mathcal{J}_{PPO}(\theta) = \mathbb{E}[\underbrace{q}_{\text{question}} \sim P(Q), o \sim \pi_{\theta_{old}}(O|q)] \frac{1}{|o|} \sum_{t=1}^{|o|} \min \left[\frac{\pi_{\theta}(o_t|q, o_{<t})}{\pi_{\theta_{old}}(o_t|q, o_{<t})} A_t, \text{clip} \left(\frac{\pi_{\theta}(o_t|q, o_{<t})}{\pi_{\theta_{old}}(o_t|q, o_{<t})}, 1 - \epsilon, 1 + \epsilon \right) A_t \right]$$

Background information about RL

- Proximal Policy Optimization (PPO)
 - Goal: improves our agent's training stability by avoiding policy updates that are too large
 - Objective

$$\mathcal{J}_{PPO}(\theta) = \mathbb{E}[\underbrace{q}_{\text{question}} \sim P(Q), \underbrace{o}_{\text{output}} \sim \pi_{\theta_{old}}(O|q)] \frac{1}{|o|} \sum_{t=1}^{|o|} \min \left[\frac{\pi_{\theta}(o_t|q, o_{<t})}{\pi_{\theta_{old}}(o_t|q, o_{<t})} A_t, \text{clip} \left(\frac{\pi_{\theta}(o_t|q, o_{<t})}{\pi_{\theta_{old}}(o_t|q, o_{<t})}, 1 - \epsilon, 1 + \epsilon \right) A_t \right]$$

Background information about RL

- Proximal Policy Optimization (PPO)
 - Goal: improves our agent's training stability by avoiding policy updates that are too large
 - Objective

$$\mathcal{J}_{PPO}(\theta) = \mathbb{E} \left[q \sim P(Q), o \sim \pi_{\theta_{old}}(O|q) \right] \frac{1}{|o|} \sum_{t=1}^{|o|} \min \left[\frac{\pi_{\theta}(o_t|q, o_{<t})}{\pi_{\theta_{old}}(o_t|q, o_{<t})} A_t, \text{clip} \left(\frac{\pi_{\theta}(o_t|q, o_{<t})}{\pi_{\theta_{old}}(o_t|q, o_{<t})}, 1 - \epsilon, 1 + \epsilon \right) A_t \right]$$

question

output

token number

Background information about RL

- Proximal Policy Optimization (PPO)
 - Goal: improves our agent's training stability by avoiding policy updates that are too large
 - Objective

$$\mathcal{J}_{PPO}(\theta) = \mathbb{E} \left[q \sim P(Q), o \sim \pi_{\theta_{old}}(O|q) \right] \frac{1}{|o|} \sum_{t=1}^{|o|} \min \left[\frac{\pi_{\theta}(o_t|q, o_{<t})}{\pi_{\theta_{old}}(o_t|q, o_{<t})} A_t, \text{clip} \left(\frac{\pi_{\theta}(o_t|q, o_{<t})}{\pi_{\theta_{old}}(o_t|q, o_{<t})}, 1 - \epsilon, 1 + \epsilon \right) A_t \right]$$

question

output

token number

policy model to be trained

Background information about RL

- Proximal Policy Optimization (PPO)
 - Goal: improves our agent's training stability by avoiding policy updates that are too large
 - Objective

$$\mathcal{J}_{PPO}(\theta) = \mathbb{E} \left[q \sim P(Q), o \sim \pi_{\theta_{old}}(O|q) \right] \frac{1}{|o|} \sum_{t=1}^{|o|} \min \left[\frac{\pi_{\theta}(o_t|q, o_{<t})}{\pi_{\theta_{old}}(o_t|q, o_{<t})} A_t, \text{clip} \left(\frac{\pi_{\theta}(o_t|q, o_{<t})}{\pi_{\theta_{old}}(o_t|q, o_{<t})}, 1 - \epsilon, 1 + \epsilon \right) A_t \right]$$

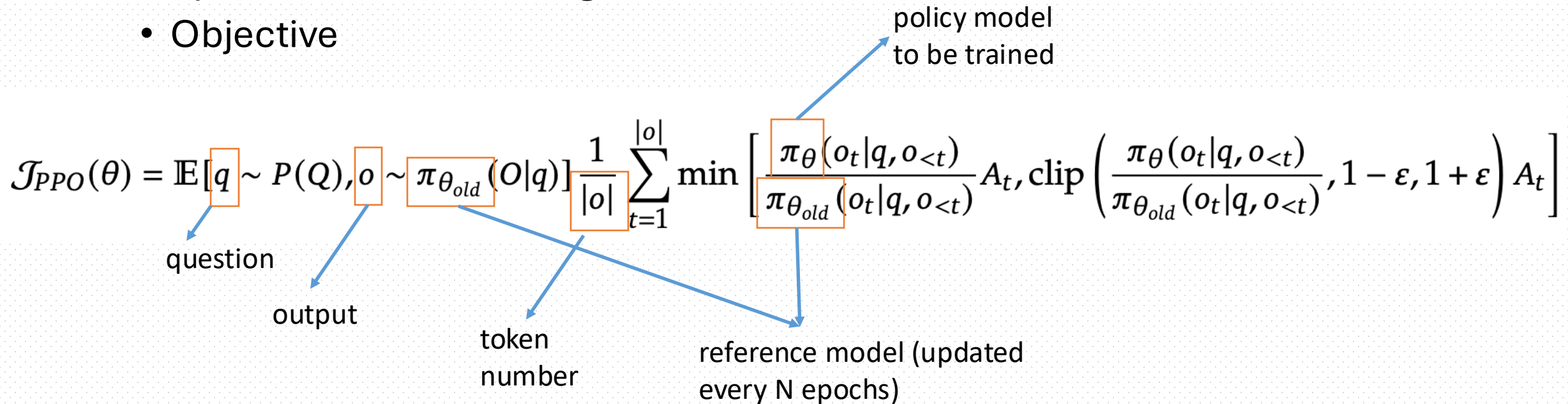
question

output

token number

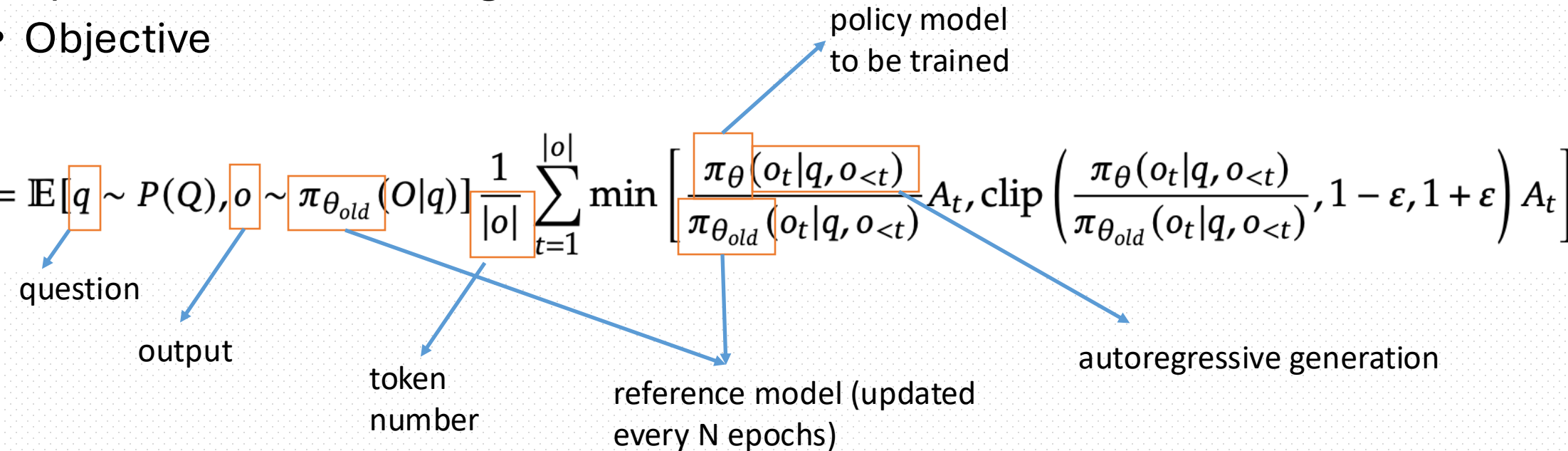
reference model (updated every N epochs)

policy model to be trained



Background information about RL

- Proximal Policy Optimization (PPO)
 - Goal: improves our agent's training stability by avoiding policy updates that are too large
 - Objective

$$\mathcal{J}_{PPO}(\theta) = \mathbb{E} \left[q \sim P(Q), o \sim \pi_{\theta_{old}}(O|q) \right] \frac{1}{|o|} \sum_{t=1}^{|o|} \min \left[\frac{\pi_{\theta}(o_t|q, o_{<t})}{\pi_{\theta_{old}}(o_t|q, o_{<t})} A_t, \text{clip} \left(\frac{\pi_{\theta}(o_t|q, o_{<t})}{\pi_{\theta_{old}}(o_t|q, o_{<t})}, 1 - \epsilon, 1 + \epsilon \right) A_t \right]$$


question

output

token number

reference model (updated every N epochs)


policy model to be trained

autoregressive generation

Background information about RL

- Proximal Policy Optimization (PPO)
 - Objective

$$\mathcal{J}_{PPO}(\theta) = \mathbb{E}[q \sim P(Q), o \sim \pi_{\theta_{old}}(O|q)] \frac{1}{|o|} \sum_{t=1}^{|o|} \min \left[\frac{\pi_{\theta}(o_t|q, o_{<t})}{\pi_{\theta_{old}}(o_t|q, o_{<t})} A_t, \text{clip} \left(\frac{\pi_{\theta}(o_t|q, o_{<t})}{\pi_{\theta_{old}}(o_t|q, o_{<t})}, 1 - \epsilon, 1 + \epsilon \right) A_t \right]$$


$$\delta_t = r_t + \gamma V^{\pi}(s_{t+1}) - V^{\pi}(s_t)$$
$$A_t = \delta_t + \gamma \lambda A_{t+1}$$

Advantage: how taking that action at a state is better compared to the average value of the state

<https://arxiv.org/abs/1506.02438>

Background information about RL

- Proximal Policy Optimization (PPO)
 - Objective

$$\mathcal{J}_{PPO}(\theta) = \mathbb{E}[q \sim P(Q), o \sim \pi_{\theta_{old}}(O|q)] \frac{1}{|o|} \sum_{t=1}^{|o|} \min \left[\frac{\pi_{\theta}(o_t|q, o_{<t})}{\pi_{\theta_{old}}(o_t|q, o_{<t})} A_t, \text{clip} \left(\frac{\pi_{\theta}(o_t|q, o_{<t})}{\pi_{\theta_{old}}(o_t|q, o_{<t})}, 1 - \epsilon, 1 + \epsilon \right) A_t \right]$$

Ratio function

Background information about RL

- Proximal Policy Optimization (PPO)
 - Objective

$$\mathcal{J}_{PPO}(\theta) = \mathbb{E}[q \sim P(Q), o \sim \pi_{\theta_{old}}(O|q)] \frac{1}{|o|} \sum_{t=1}^{|o|} \min \left[\frac{\pi_{\theta}(o_t|q, o_{<t})}{\pi_{\theta_{old}}(o_t|q, o_{<t})} A_t, \text{clip} \left(\frac{\pi_{\theta}(o_t|q, o_{<t})}{\pi_{\theta_{old}}(o_t|q, o_{<t})}, 1 - \epsilon, 1 + \epsilon \right) A_t \right]$$

clip: clipped the ratio between $[1 - \epsilon, 1 + \epsilon]$

$p_t(\theta) > 0$	A_t	Return Value of \min	Objective is Clipped	Sign of Objective	Gradient
$p_t(\theta) \in [1 - \epsilon, 1 + \epsilon]$	+	$p_t(\theta) A_t$	no	+	✓
$p_t(\theta) \in [1 - \epsilon, 1 + \epsilon]$	-	$p_t(\theta) A_t$	no	-	✓
$p_t(\theta) < 1 - \epsilon$	+	$p_t(\theta) A_t$	no	+	✓
$p_t(\theta) < 1 - \epsilon$	-	$(1 - \epsilon) A_t$	yes	-	0
$p_t(\theta) > 1 + \epsilon$	+	$(1 + \epsilon) A_t$	yes	+	0
$p_t(\theta) > 1 + \epsilon$	-	$p_t(\theta) A_t$	no	-	✓

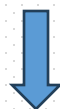
Background information about RL



- Three popular RL framework
 - TRL: <https://github.com/huggingface/trl>
 - OpenRLHF: <https://github.com/OpenRLHF/OpenRLHF>
 - veRL: <https://github.com/volcengine/verl>
 - SOTA
 - <https://arxiv.org/pdf/2409.19256> (EuroSys 2025)

- GRPO

$$\mathcal{J}_{PPO}(\theta) = \mathbb{E}[q \sim P(Q), o \sim \pi_{\theta_{old}}(O|q)] \frac{1}{|o|} \sum_{t=1}^{|o|} \min \left[\frac{\pi_{\theta}(o_t|q, o_{<t})}{\pi_{\theta_{old}}(o_t|q, o_{<t})} A_t, \text{clip} \left(\frac{\pi_{\theta}(o_t|q, o_{<t})}{\pi_{\theta_{old}}(o_t|q, o_{<t})}, 1 - \epsilon, 1 + \epsilon \right) A_t \right]$$



$$\begin{aligned} \delta_t &= r_t + \gamma V^{\pi}(s_{t+1}) - V^{\pi}(s_t) \\ A_t &= \delta_t + \gamma \lambda A_{t+1} \end{aligned}$$

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^G \sim \pi_{\theta_{old}}(O|q)]$$

$$\frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \left\{ \min \left[\frac{\pi_{\theta}(o_{i,t}|q, o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q, o_{i,<t})} \hat{A}_{i,t}, \text{clip} \left(\frac{\pi_{\theta}(o_{i,t}|q, o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q, o_{i,<t})}, 1 - \epsilon, 1 + \epsilon \right) \hat{A}_{i,t} \right] - \beta \mathbb{D}_{KL} [\pi_{\theta} || \pi_{ref}] \right\}$$

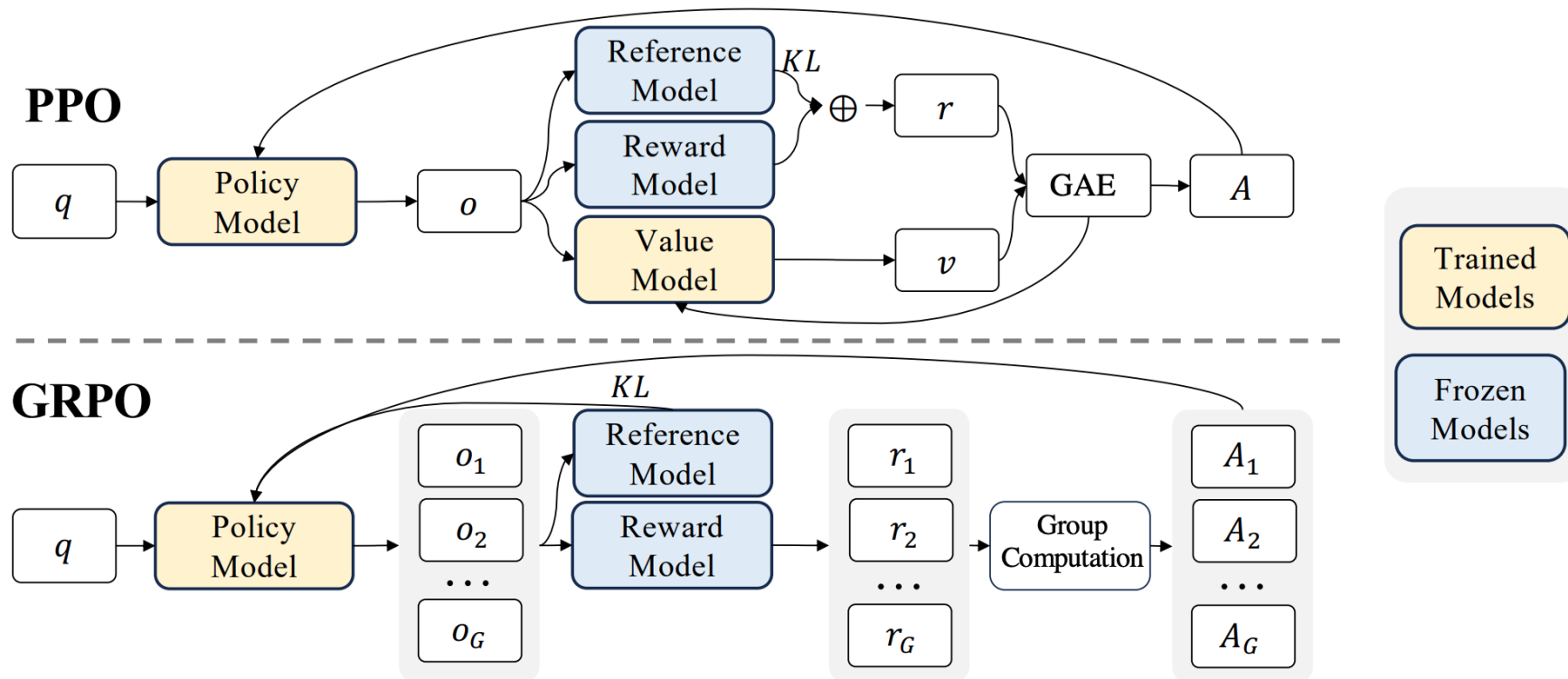
$$\mathbf{R} = \{ \{r_1^{index(1)}, \dots, r_1^{index(K_1)}\}, \dots, \{r_G^{index(1)}, \dots, r_G^{index(K_G)}\} \}$$

$$\tilde{r}_i^{index(j)} = \frac{r_i^{index(j)} - \text{mean}(\mathbf{R})}{\text{std}(\mathbf{R})}$$

$$\hat{A}_{i,t} = \sum_{index(j) \geq t} \tilde{r}_i^{index(j)}$$

No value model!

- GRPO

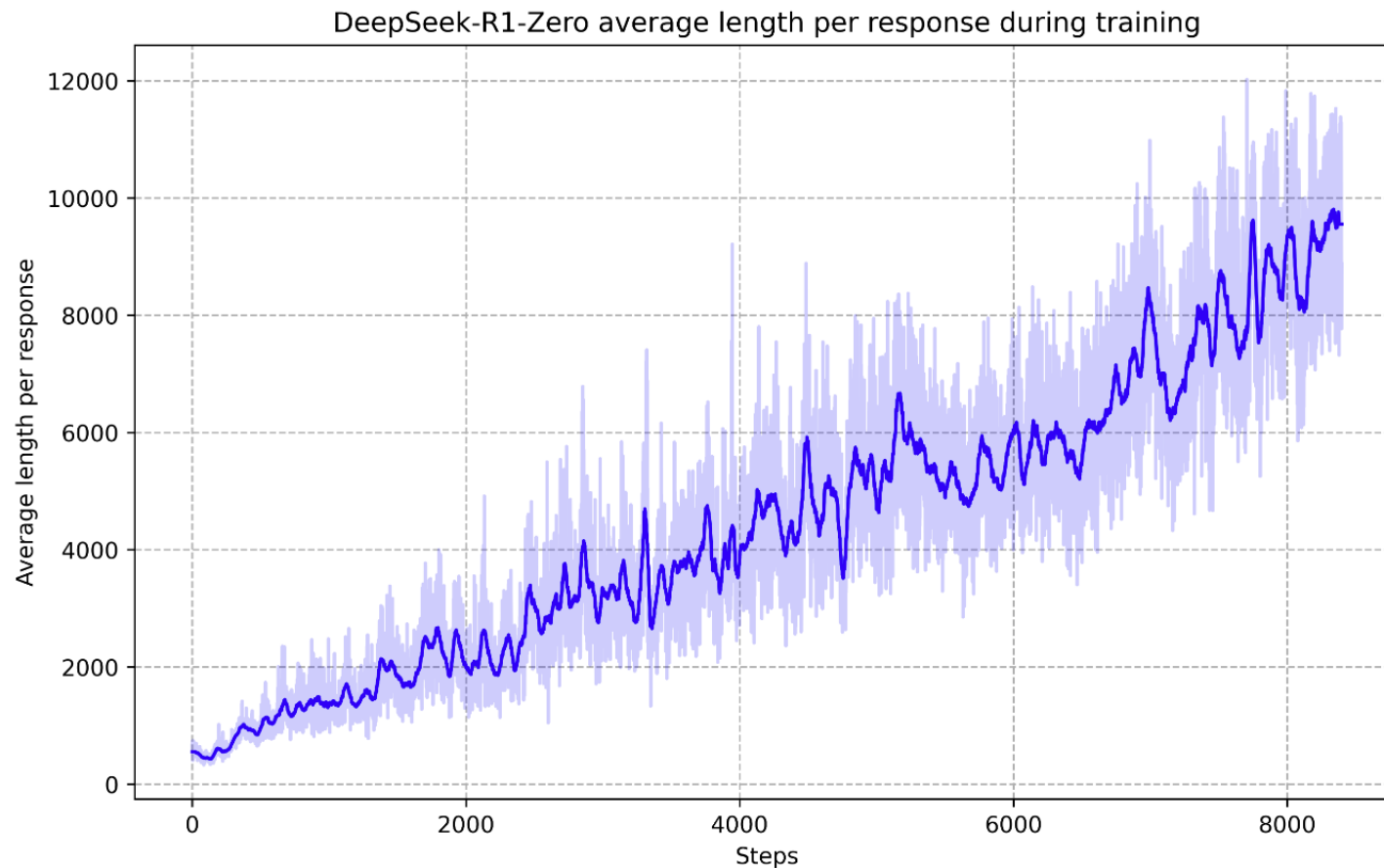


<https://arxiv.org/pdf/2402.03300>

- GRPO
 - Disadvantages of value model
 - Substantial memory and computational burden
 - Hard to be accurate for every token

DeepSeek-R1-Zero

- Reward design
 - Rule-based reward
 - Accuracy reward
 - Format reward
- Aha moment
 - Average response length goes up during training
 - Model learns to rethink



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 ...

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

Rearrange to isolate the inner square root term:

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

...

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.

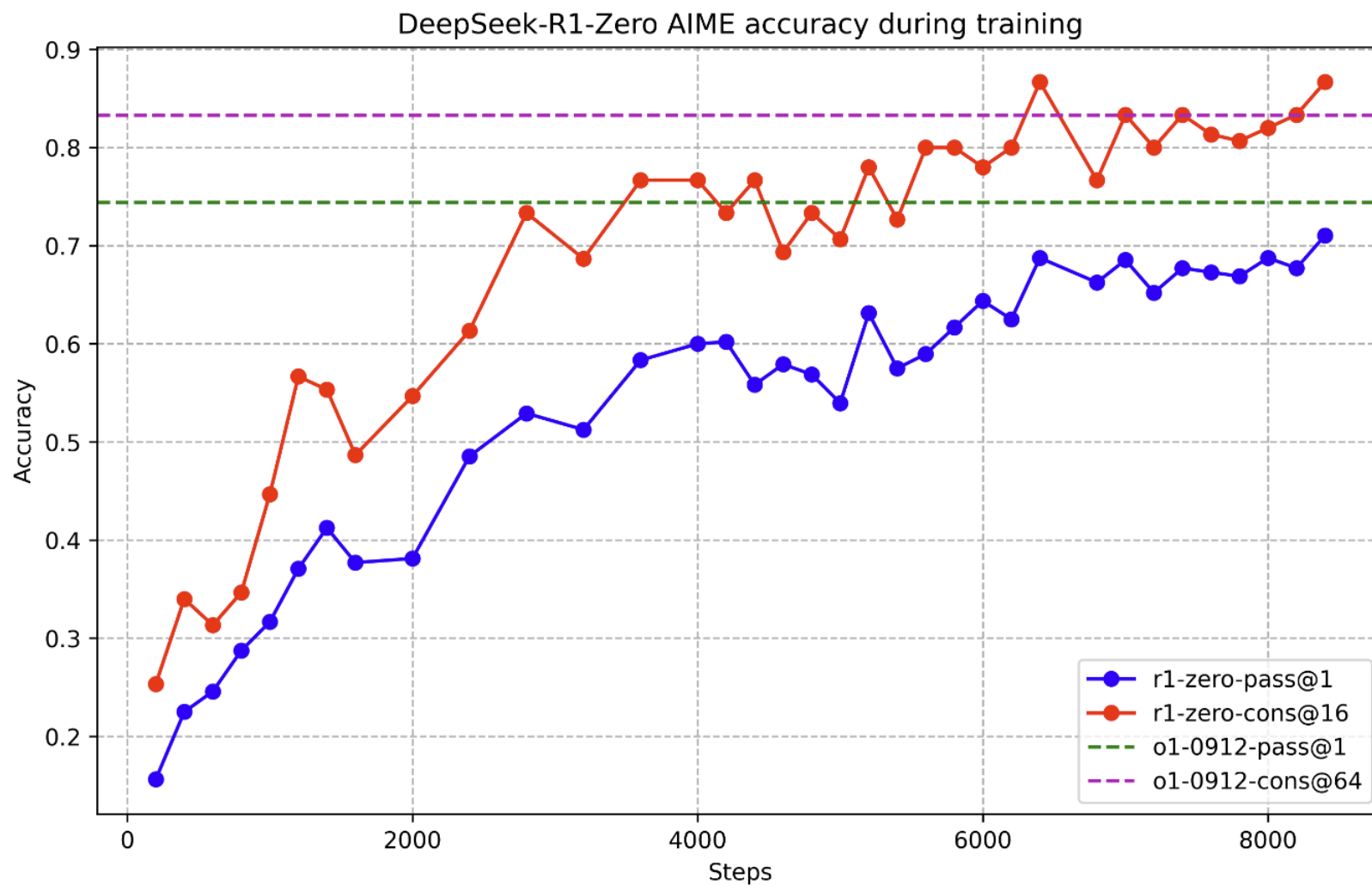
DeepSeek-R1-Zero

- Evaluation

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

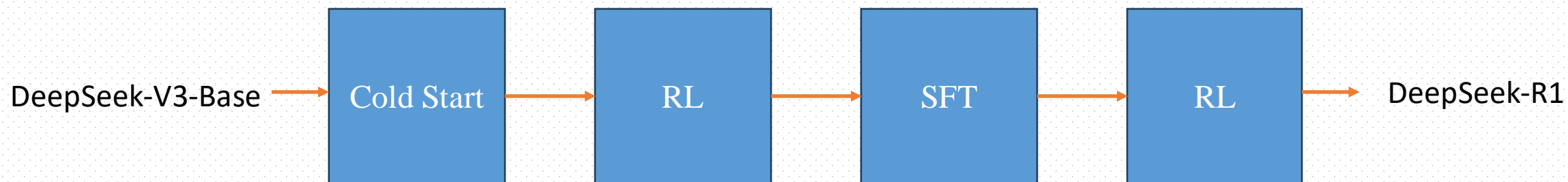
Pure RL works!

- Evaluation

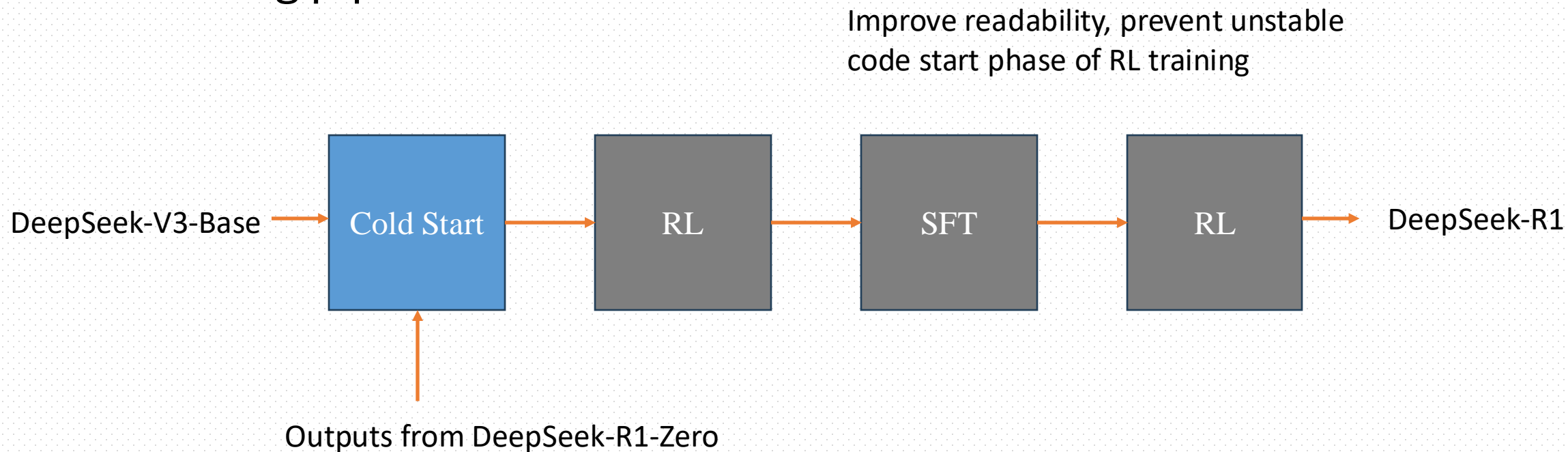


DeepSeek-R1

- Training pipeline

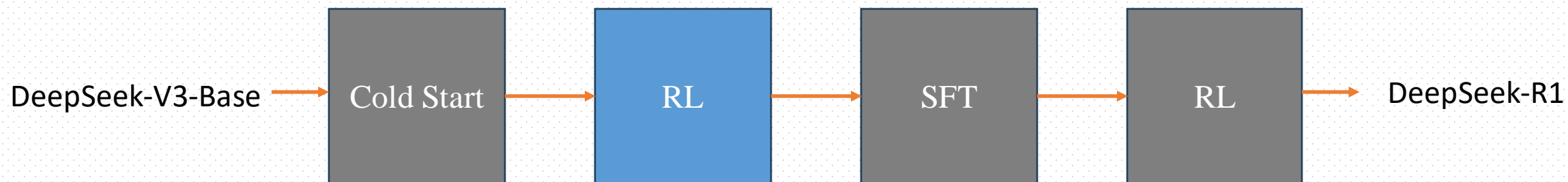


- Training pipeline



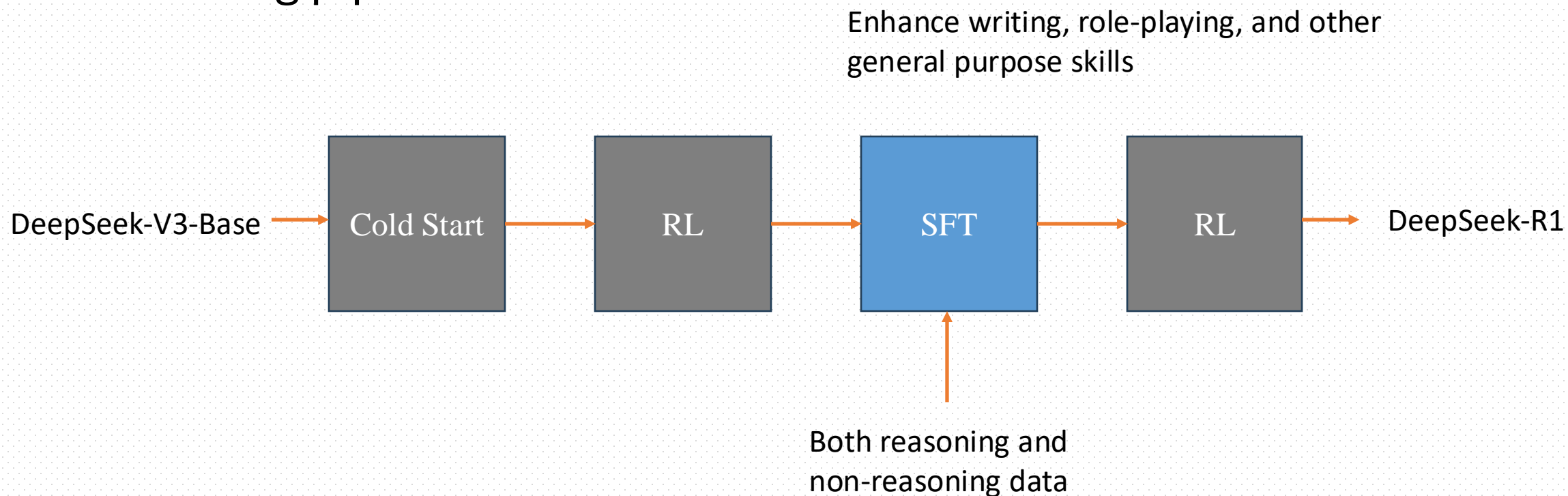
- Training pipeline

Focus on enhancing reasoning capabilities
(coding, mathematics, etc.)

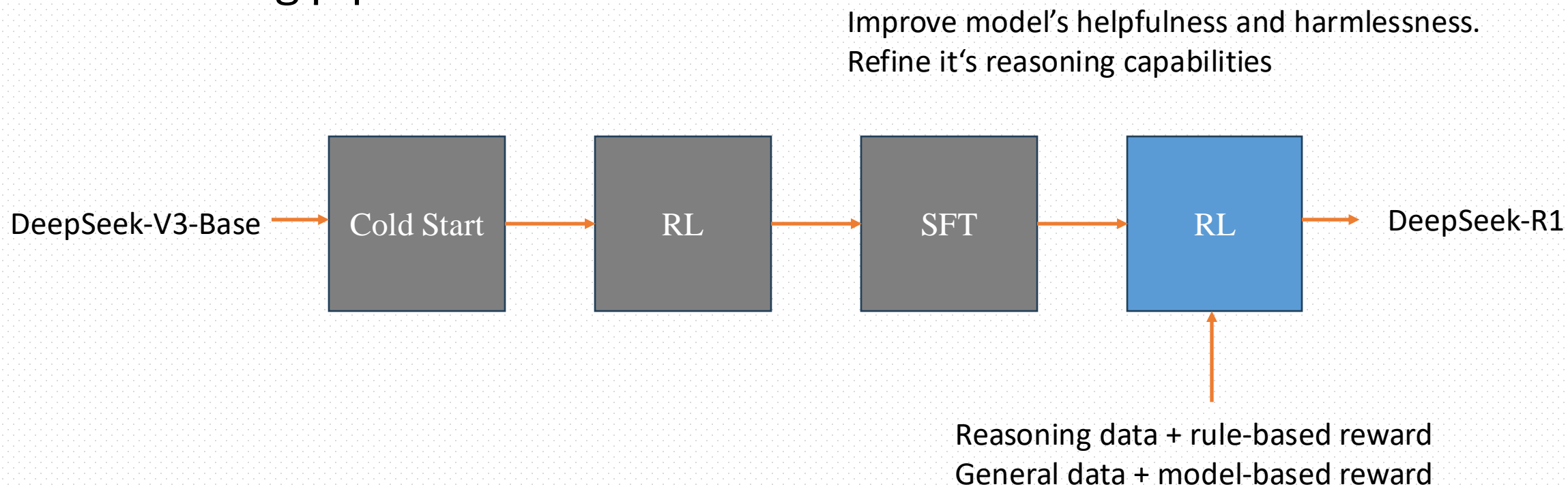


Same training process as
DeepSeek-R1-Zero

- Training pipeline



- Training pipeline



DeepSeek-R1

- Distillation
 - Use samples curated with DeepSeek-R1 to fine-tune small models

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

- Discussion
 - Distillation v.s. Reinforcement Learning

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

- Discussion
 - Unsuccessful attempts
 - Process Reward Model (PRM)
 - Hard to define step in general reasoning
 - Hard to evaluate intermediate step
 - Model-base PRM leads to reward hacking and consumes training resources
 - Monte Carlo Tree Search (MCTS)
 - Huge search space make it hard to scale up
 - Training a good value model to guide this search process is also difficult

- Reinforcement learning is a powerful tool
 - Significant discovery: model autonomously acquires test-time scaling capabilities during through the reinforcement learning process
 - GRPO optimization algorithm: efficient, memory-saving
 - Reward design: accuracy + format (rule-based ORM)
 - Both algorithm and infra matters
- Distillation is a cost-effective yet promising method

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

Author: DeepSeek-AI

Thank you for listening!

