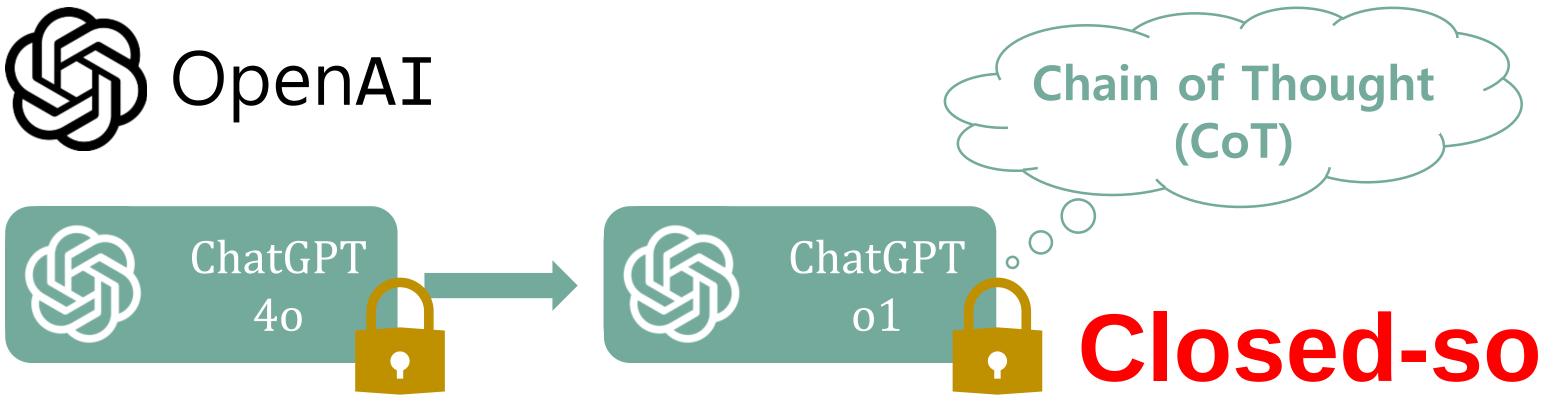
deepseek-R1

Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

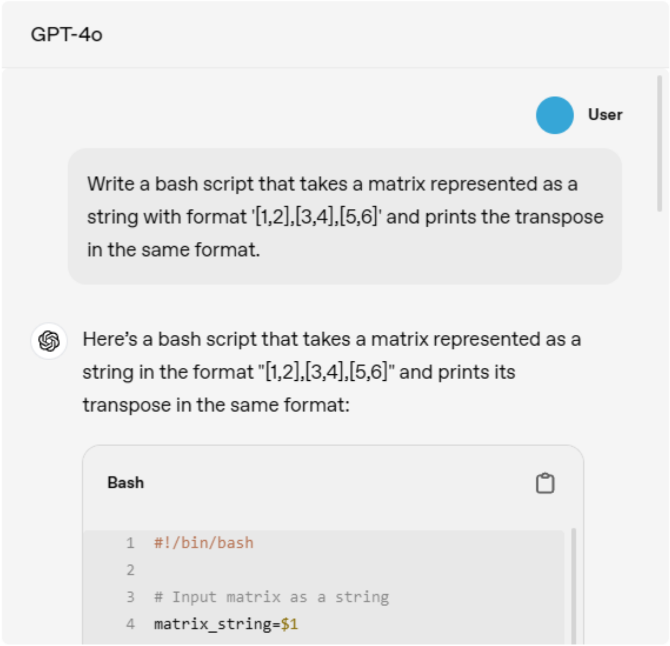
서울대학교 IDEA 연구실
박사연구생 박세현

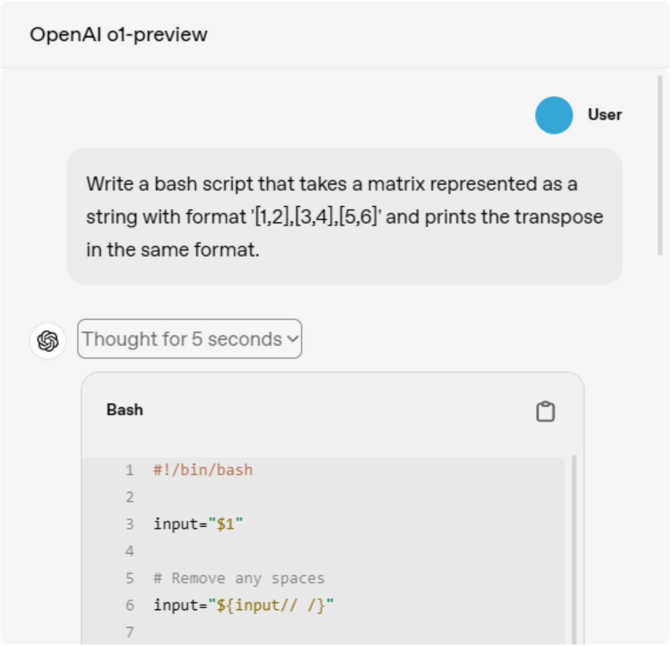
OUTLINE

01. Introduction



Closed-source!





01. Introduction

deepseek-R1

deepseek

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.

Q. What are the differences of DeepSeek-R1?

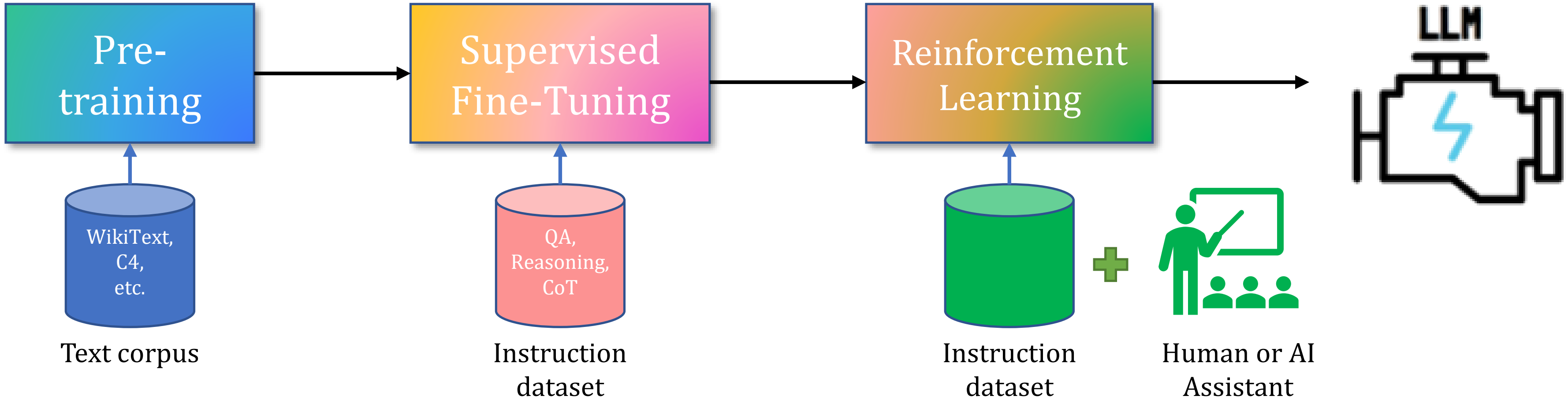
1. **Low-Cost Development of High-Performance AI Models:**
   - DeepSeek has stated that it spent approximately $6 million to develop the "R1" model.
   - This is significantly cheaper compared to companies like OpenAI and Google, which have invested billions of dollars.
2. **Open-Source Strategy:**
   - DeepSeek has made "R1" open-source.
   - DeepSeek-R1 is 720GB, but "**Unsloth**" released a 131GB version with similar performance.

unsloth

02. DeepSeek-R1-Zero

▶ Training Process of Large Language Models (LLMs)

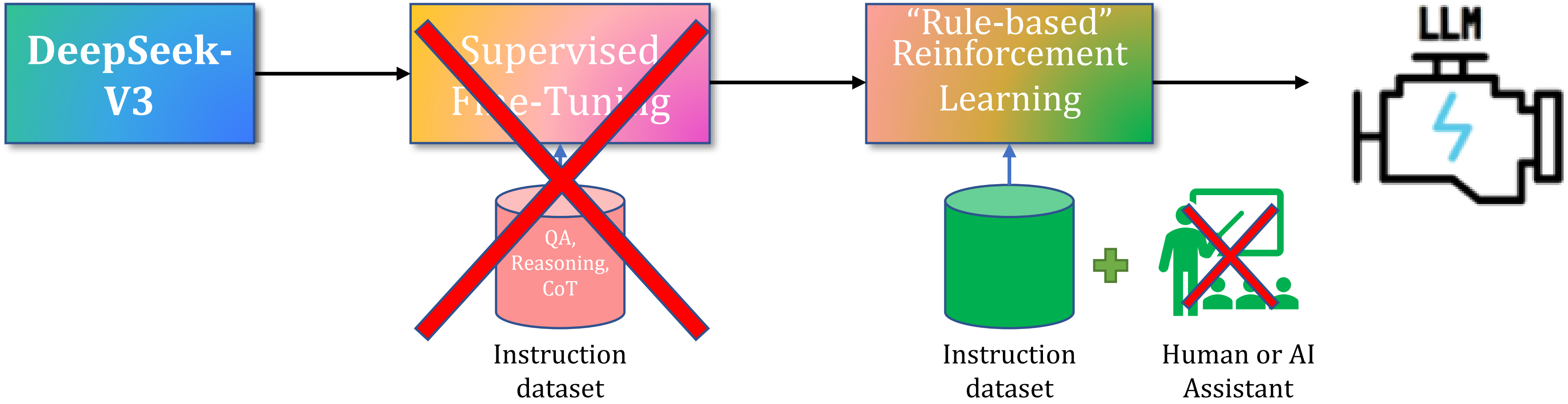
1. **Pre-training:**
   - The model is pre-trained on a vast amount of text and code data to acquire general knowledge.
   - It learns sentence prediction but still struggles to follow explicit human instructions.
2. **Supervised Fine-Tuning (SFT)**
   - The model is fine-tuned using a **(text-label) dataset** to enhance its ability to follow instructions.
3. **Reinforcement Learning with Feedback (RLHF, RLAIF)**
   - The model is further improved using **human feedback (RLHF) or AI-generated feedback (RLAIF).**



02. DeepSeek-R1-Zero

▶ DeepSeek-R1-Zero's Approach

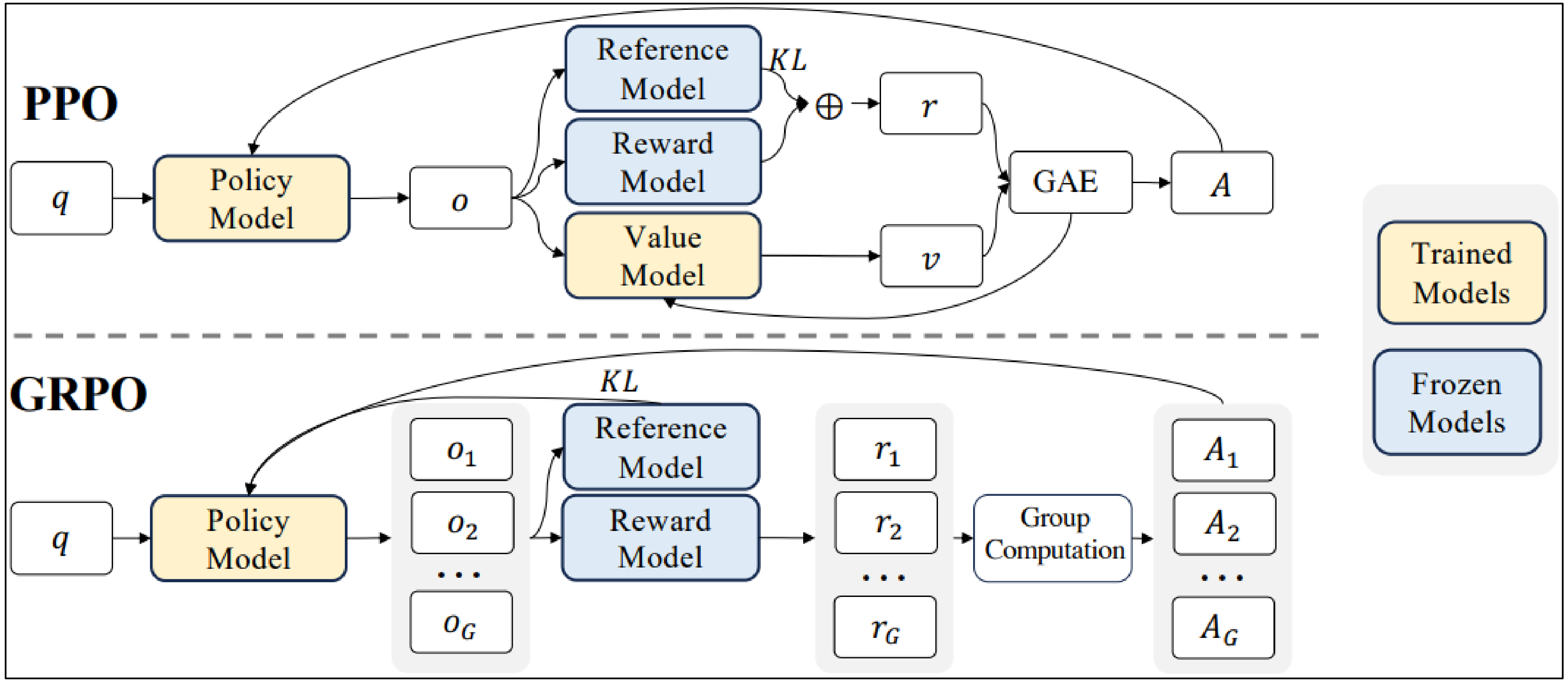
- Conducting experiments to **enhance AI's reasoning ability using pure reinforcement learning (RL) without supervised learning**.
- Utilizing the **DeepSeek-V3-Base model (parameters : 671 billion)** and applying the **GRPO technique (a method for optimizing relative policies within a specific group)**.
- For reinforcement learning, they adopt a **rule-based reward system. (Accuracy, Format)**



02. DeepSeek-R1-Zero

▶ Group Relative Policy Optimization (GRPO)

- In order to save the training costs of RL, Deepseek-R1-Zero adopt GRPO, which foregoes the critic model that is typically the same size as the policy model, and estimates the baseline from group scores instead.



Ref : Shao, Zhihong, et al. "Deepseekmath: Pushing the limits of mathematical reasoning in open language models." *arXiv preprint arXiv:2402.03300* (2024).

02. DeepSeek-R1-Zero

▶ Reward Modeling

- They does not use a neural model to generate rewards, simplifies and reduces the cost of the training process, making it feasible at a large scale
- To train DeepSeek-R1-Zero, they adopt a **rule-vased reward system** that mainly consists of two types of rewards:

1. **Accuracy rewards:**
    - The accuracy reward model evaluates whether the response is correct.
    - For instance, in math problems with deterministic results, we can reliably check if the final answer provided by the model is correct.
    - For code problems with predefined test cases, a compiler generates feedback based on the test cases.
2. **Format rewards**
    - They employ a format reward model that enforces the model to put its thinking process between ‘<think>’ and ‘</think>’ tags.

02. DeepSeek-R1-Zero

▶ Training 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:

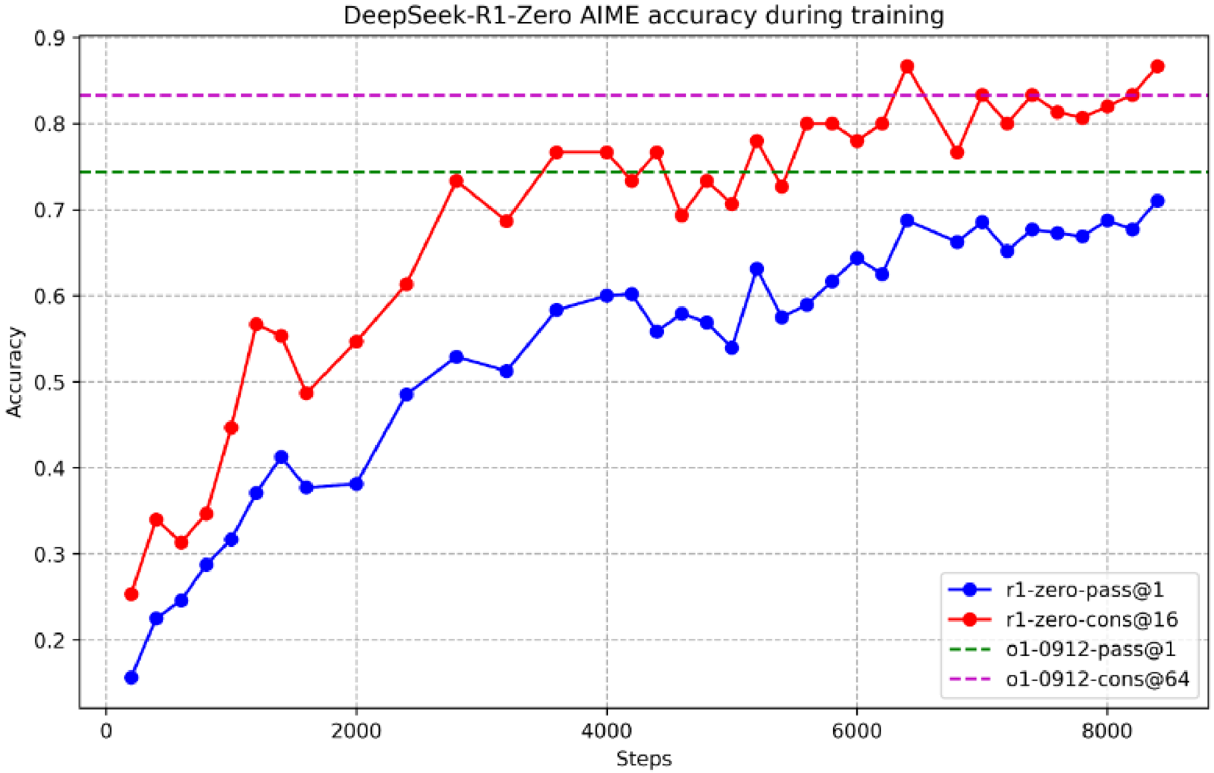
Table 1 | Template for DeepSeek-R1-Zero. prompt will be replaced with the specific reasoning question during training.

Ref : Guo, Daya, et al. "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning." *arXiv preprint arXiv:2501.12948* (2025).

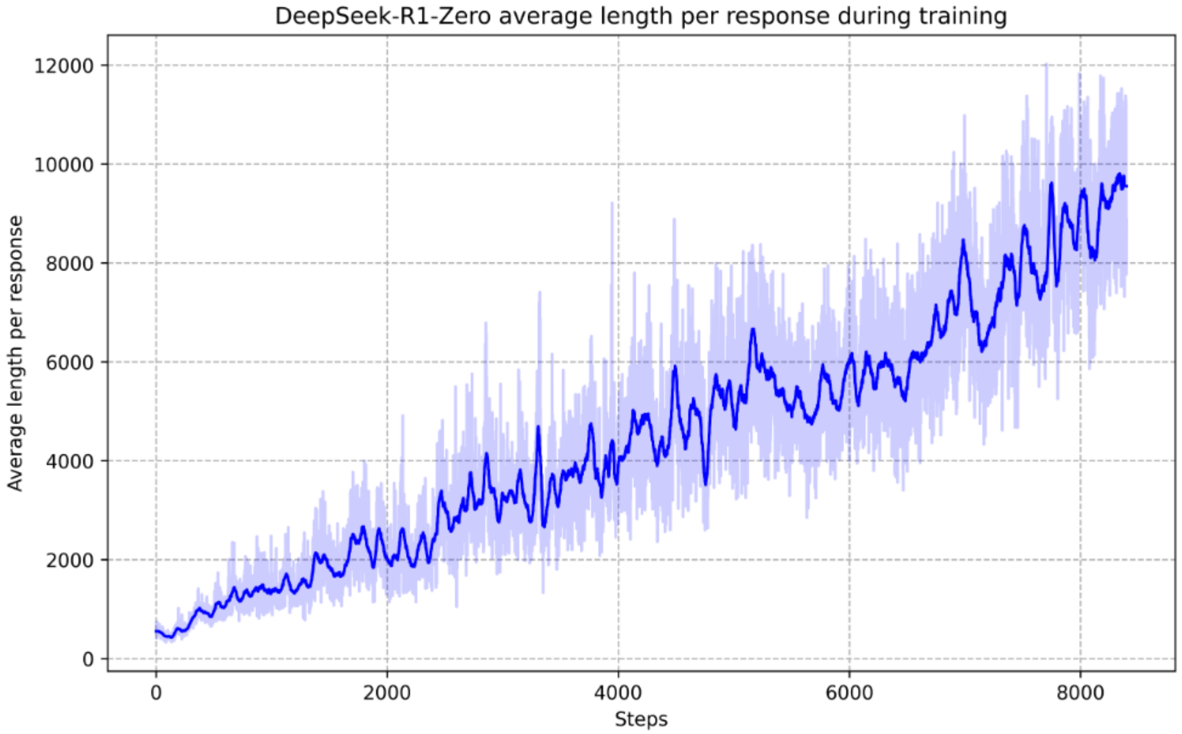
02. DeepSeek-R1-Zero

▶ Performance and Aha Moment of DeepSeek-R1-Zero

〈AIME accuracy of DeepSeek-R1-Zero during training〉



〈The average response length of DeepSeek-R1-Zero〉



Ref : Guo, Daya, et al. "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning." *arXiv preprint arXiv:2501.12948* (2025).

02. DeepSeek-R1-Zero

▶ Performance and Aha Moment of DeepSeek-R1-Zero

〈Aha Moment of DeepSeek-R1-Zero〉

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: $\cdots$
$\ldots$

Ref : Guo, Daya, et al. "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning." *arXiv preprint arXiv:2501.12948* (2025).

02. DeepSeek-R1-Zero

▶ Drawback of DeepSeek-R1-Zero

- There are two main drawbacks of DeepSeek-R1-Zero

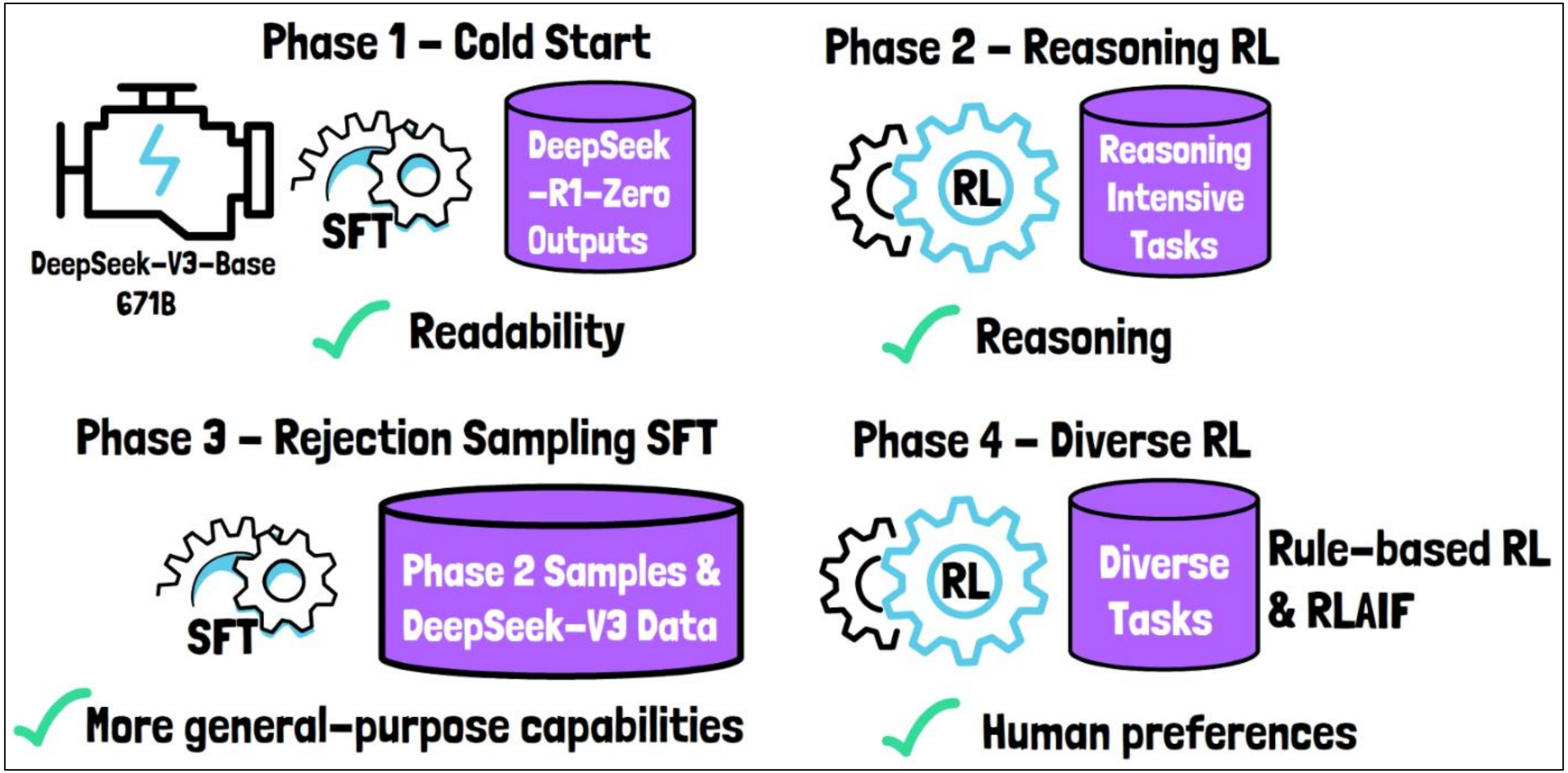
  1. **Readability Issue:**
     - DeepSeek-R1-Zero's outputs often suffer from poor readability.
  2. **Language Consistency**
     - It frequently mixes languages within a single response.

- To make reasoning processes more readable and share them with the open community, they explore **DeepSeek-R1**, a method that utilizes RL with human-friendly cold-start data.

deepseek-R1-Zero → deepseek-R1

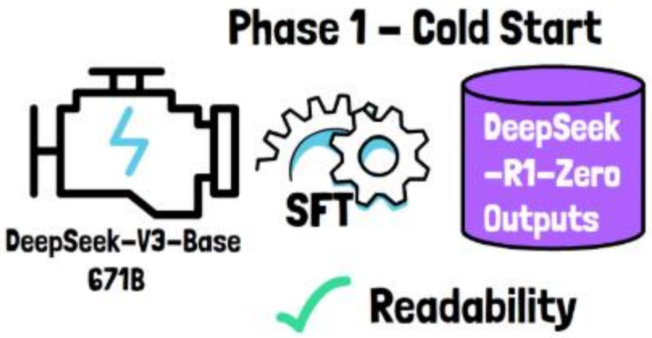
03. DeepSeek-R1

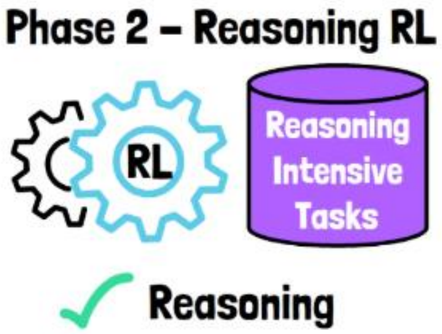
▶ Training Pipeline of DeepSeek-R1



Ref : AI Papers Academy. (2025, January 25). *Deepseek-R1 paper explained - a new RL LLMS era in ai?* https://aipapersacademy.com/deepseek-r1/

03. DeepSeek-R1





Reasoning ability & Language consistency

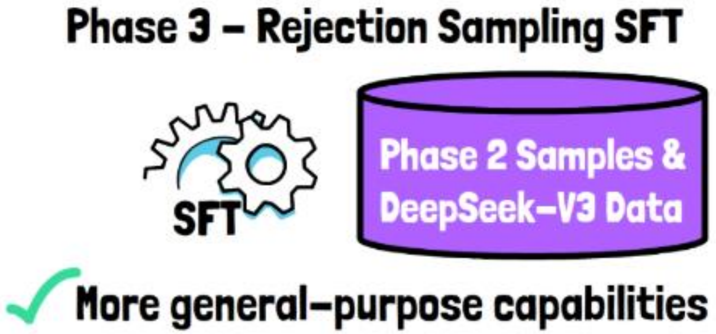
▶ Phase 1 – Cold Start

- Starting with the pre-trained model DeepSeek-V3-Base.
- Collecting a small amount of long CoT data from <u>DeepSeek-R1-Zero</u> to fine-tune the model.
- Supervised fine-tuning DeepSeek-V3-Base with this data.

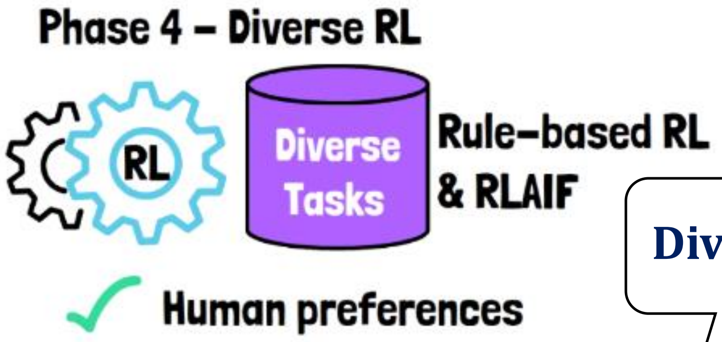
▶ Phase 2 – Reasoning RL

- Appling the same large-scale RL training process as employed in DeepSeek-R1-Zero.
- To mitigate the issue of language mixing, They add a language consistency reward.
- Task: (well-defined problems with clear solutions)
  - ➢ coding
  - ➢ mathematics
  - ➢ science
  - ➢ logic reasoning

03. DeepSeek-R1



Readability



Diverse task ability

▶ Phase 3 – Rejection Sampling SFT

- Generating many samples using the model checkpoint from phase 2.
- Using DeepSeek-V3, unreadable samples are rejected, and only readable ones are kept.
- Additionally, Some of DeepSeek-V3's training data is also included in this phase.
- Supervised fine-tuning on this dataset. ※ 800k samples

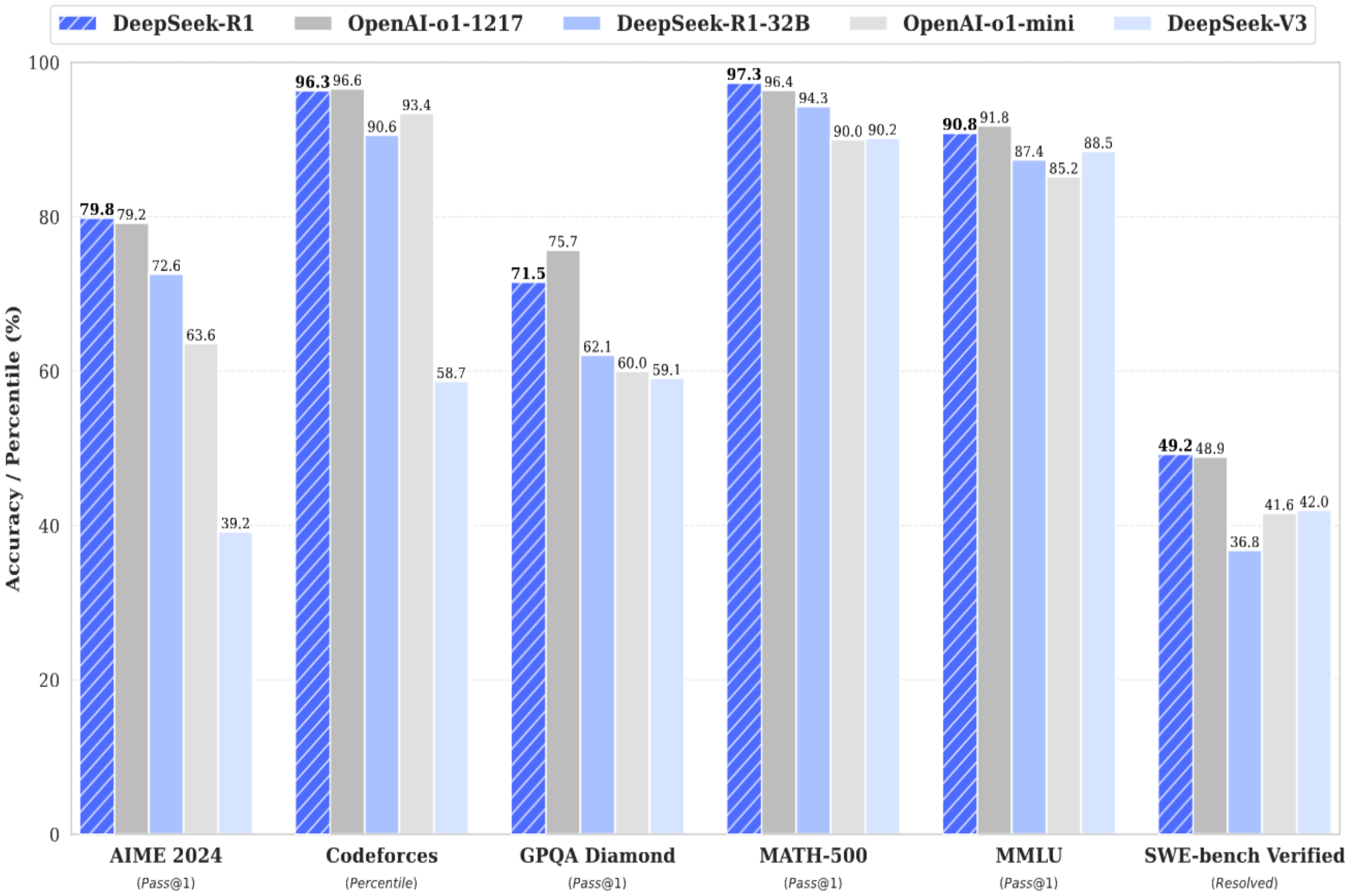
▶ Phase 4 – Diverse RL

- The final phase includes diverse tasks.

1. Reasoning data
   - ➢ Rule-based rewards are utilized for tasks that allow that, such as math.
2. General data
   - ➢ DeepSeek-V3 provides feedback to align the model with human preferences.

04. Experiments

▶ Benchmark performance of DeepSeek-R1



- AIME 2024:
  - ▷ A high-level math competition for advanced problem-solving
- Codeforces
  - ▷ An online platform for algorithmic coding contests
- GPQA Diamond
  - ▷ A question-answering test requiring deep reasoning
- MATH-500
  - ▷ A rigorous math set for theoretical and computational assessment
- MMLU
  - ▷ A broad benchmark measuring language understanding across subjects
- SWE-bench Verified
  - ▷ A software engineering measure for verified code quality

End

Appendix

▶ Algorithm for GRPO

Algorithm 1 Iterative Group Relative Policy Optimization

Input initial policy model $\pi_{\theta_{\text{init}}}$; reward models r_φ; task prompts $\mathcal{D}$; hyperparameters ε, β, μ

1: policy model $\pi_\theta \leftarrow \pi_{\theta_{\text{init}}}$
2: **for** iteration = 1, . . . , I **do**
3: reference model $\pi_{ref} \leftarrow \pi_\theta$
4: **for** step = 1, . . . , M **do**
5: Sample a batch $\mathcal{D}_b$ from $\mathcal{D}$
6: Update the old policy model $\pi_{\theta_{old}} \leftarrow \pi_\theta$
7: Sample G outputs $\{o_i\}_{i=1}^G \sim \pi_{\theta_{old}}(\cdot \mid q)$ for each question $q \in \mathcal{D}_b$
8: Compute rewards $\{r_i\}_{i=1}^G$ for each sampled output o_i by running r_φ
9: Compute $\hat{A}_{i,t}$ for the t-th token of o_i through group relative advantage estimation.
10: **for** GRPO iteration = 1, . . . , μ **do**
11: Update the policy model π_θ by maximizing the GRPO objective (Equation 21)
12: Update r_φ through continuous training using a replay mechanism.

Output π_θ

Ref : Shao, Zhihong, et al. "Deepseekmath: Pushing the limits of mathematical reasoning in open language models." *arXiv preprint arXiv:2402.03300* (2024).