

DeepSeek-R1

Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

기여: 조해창(발표자), 김보성, 임다예슬, 정현석, 이예서, 박주혜,
황예진, 김기환, 박민건, 안가영



DeepSeek-R1

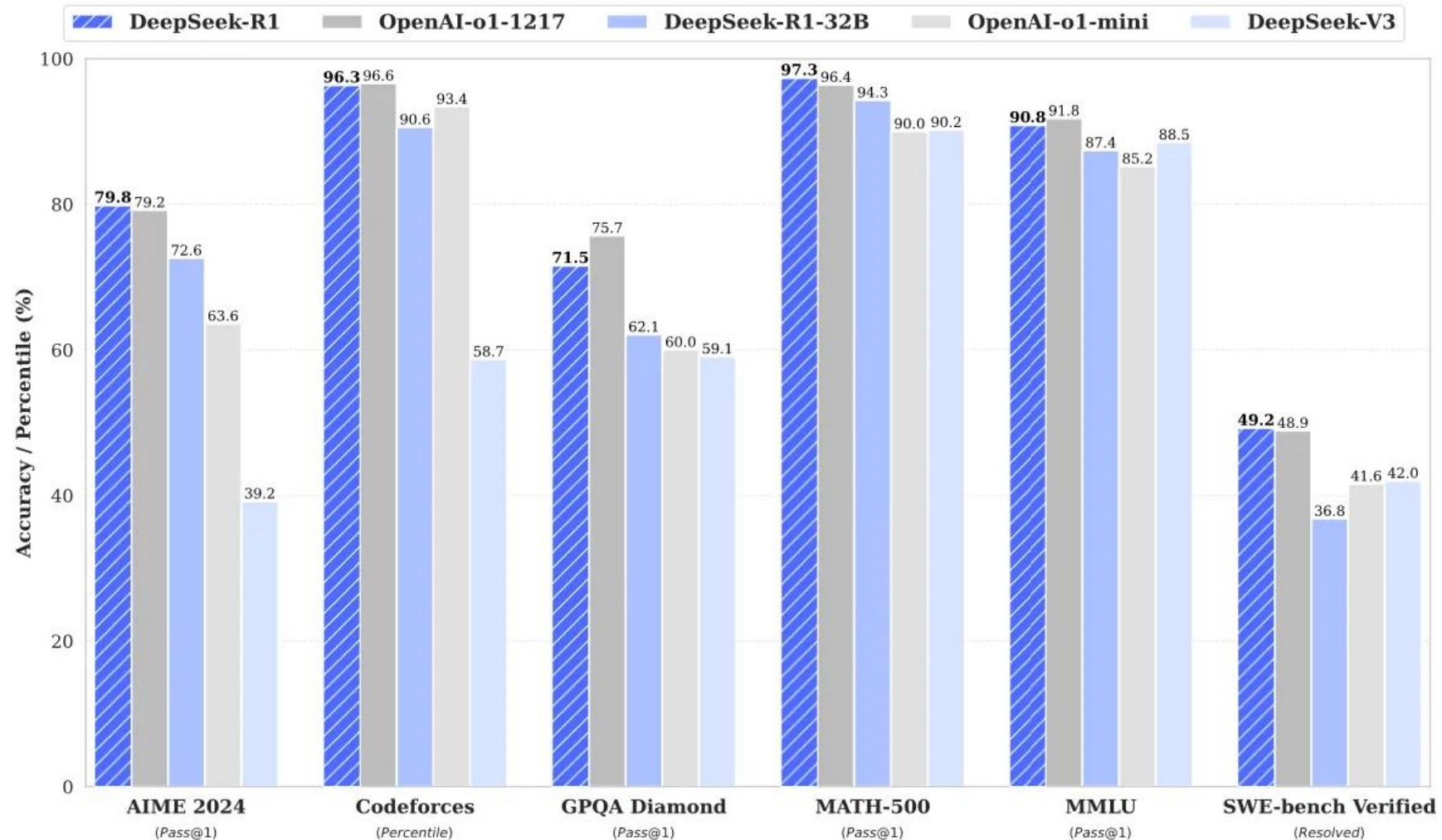


Figure 1 | Benchmark performance of DeepSeek-R1.

주요 기여점

- 언어모델에 **순수 강화학습**을 적용할 방법론 제시
- Test-time Scaling을 모델 학습으로 구현할 방법론 제시
- 뛰어난 추론능력이 KD(knowledge distillation)모델에서도 유지됨을 증명
- 상용화 모델과 비교해도 경쟁력 있는 모델의 오픈소스화

LM + size up -> LLM

- 좋은 학습 데이터
- 많은 데이터의 표현을 잘 학습할 수 있는 큰 모델
- 여러가지 테스크를 동시에 처리할 수 있도록 Instruction tuning
- 사용자 친화 동작 학습

LLM + Alignment training

- RLHF
- Human preference aware training
- Deepmind(Sparrow), OpenAI(InstuctGPT), Anthropic(Claude)
- PPO, DPO...
- 사람이 보다 선호할 답변을 만듦

• TRL ▾

Q Search documentation

⌘K

V0.15.2 ▾

EN ▾



12,120

API

TRAINERS

AlignProp

BCO

CPO

DDPO

DPO

Online DPO

GKD

GRPO

KTO

Nash-MD

ORPO

PPO

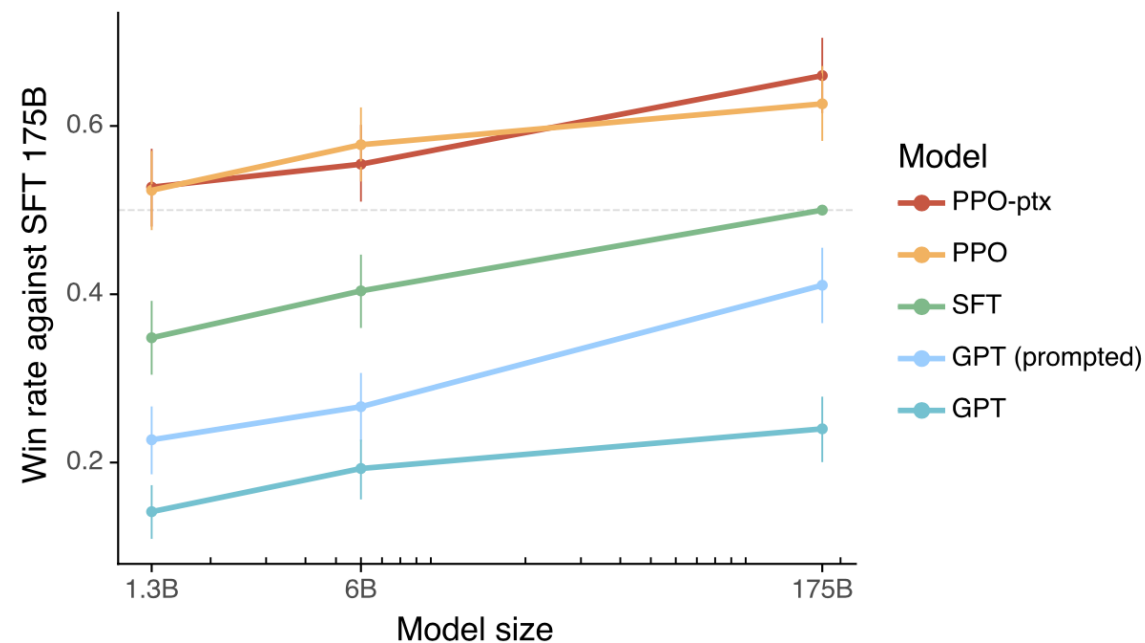
PRM

Reward

RLCO

LLM + Alignment training

- RLHF
- Human preference aware training
- Deepmind(Sparrow), OpenAI(InstuctGPT), Anthropic(Claude)
- PPO, DPO...
- 사람이 보다 선호할 답변을 만듦



TRL

Search documentation

K

V0.15.2

EN

12,120

API

TRAINERS

AlignProp

BCO

CPO

RLHF(Reinforcement Learning from Human Feedback)

- 유사 강화 학습
- 강화학습: Agent – Environment 의 상호작용에서 보상을 극대화하는 알고리즘
- 완전한 보상모델을 얻을 수 없기 때문에, 언어모델로 부터 보상값을 얻을 수 있도록 설계
- 사람의 Preference 데이터 labeling이 필요함

LLM + RL

- 언어 테스트는
정해진 정답이 있는 테스트(감정분류 등)는 단순 FT
정해진 정답이 없는 테스트(번역, 요약)는 보상값 설계가 어려움
- 강화학습으로 학습하기에 적절치 않음.
- -> RLHF
- 수학문제 처럼 정답이 있는 문제는 가능하지 않을까?

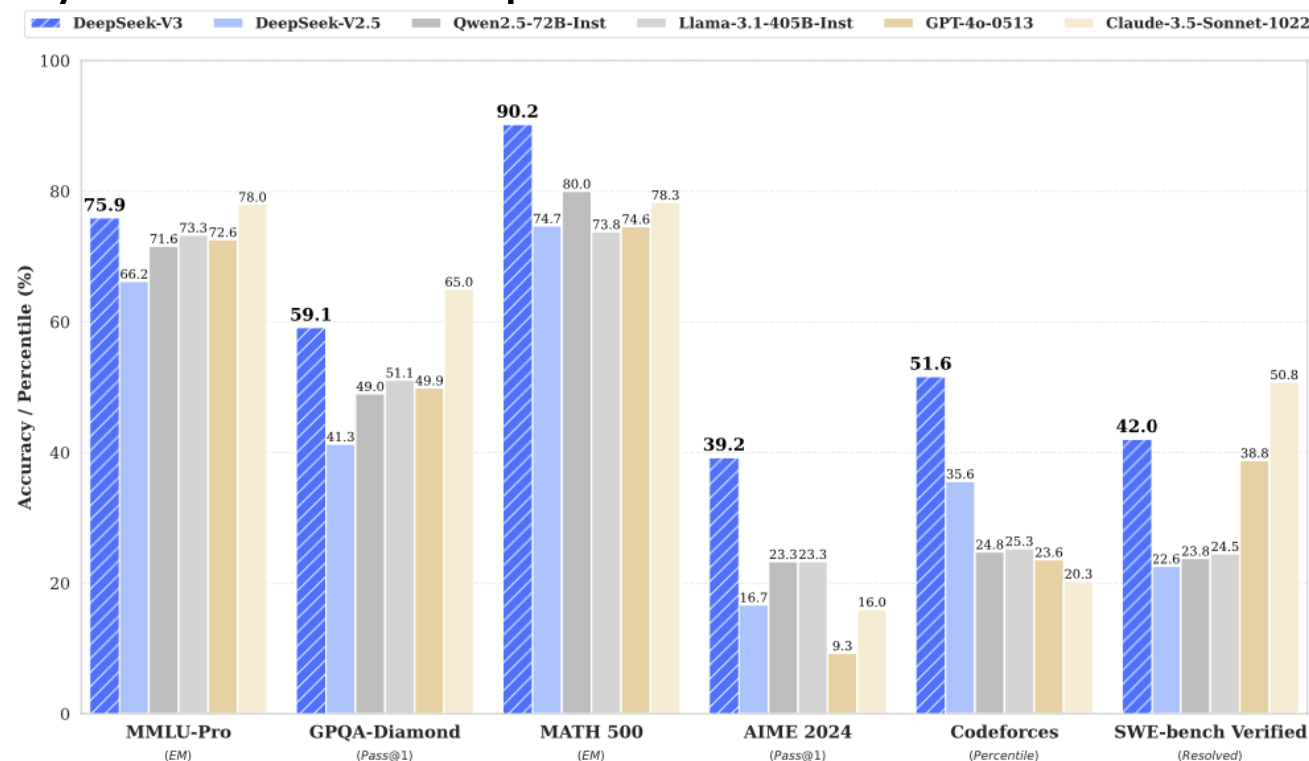
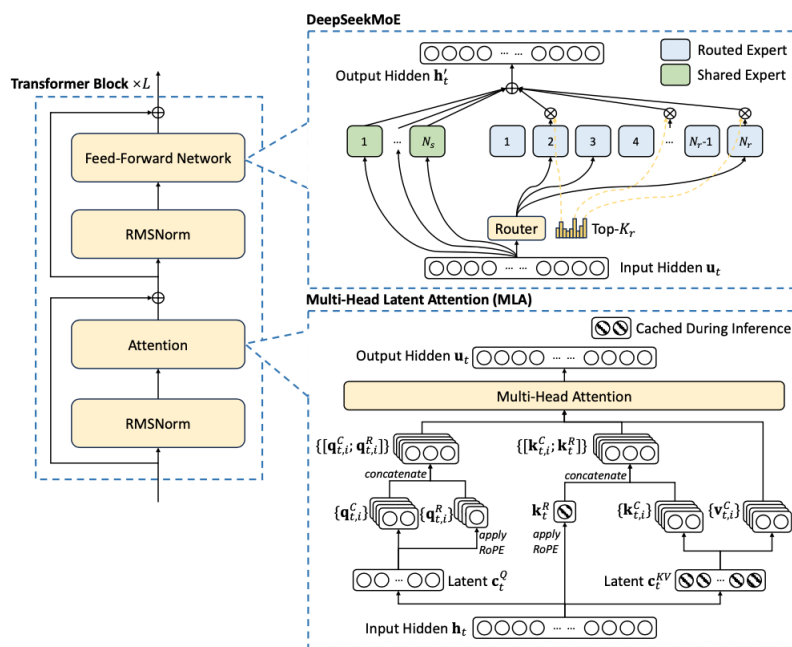
Q & A

DeepSeek

- DeepSeek-V3
- DeepSeekMath - GRPO
- DeepSeek-R1-zero
- DeepSeek-R1

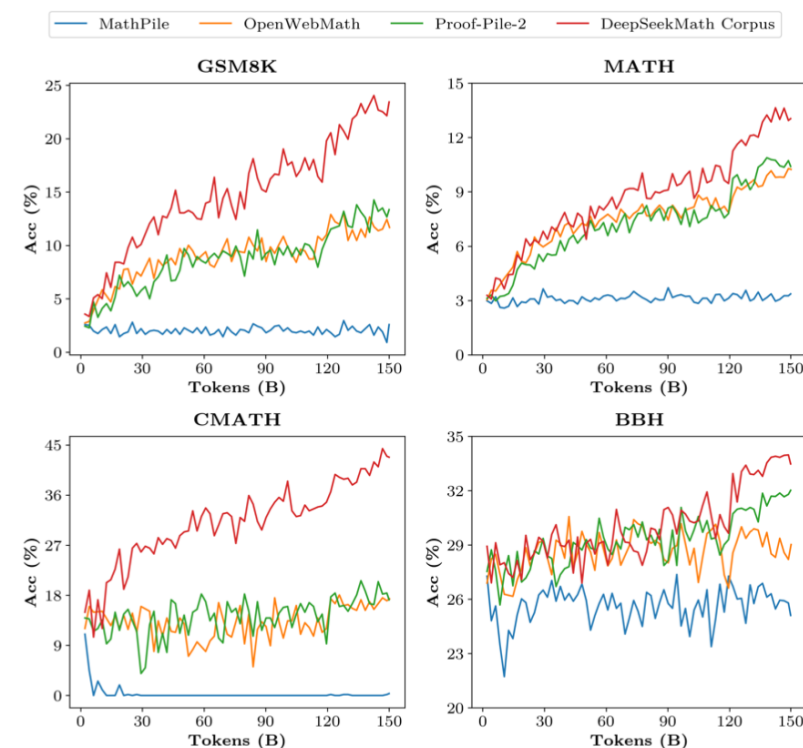
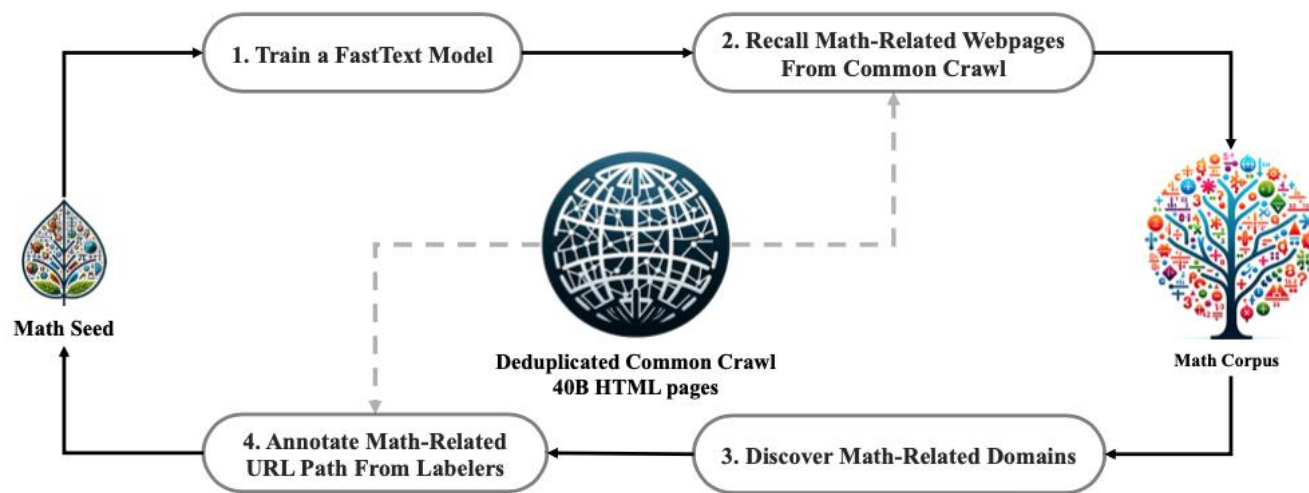
DeepSeek-V3

- 671B pretrained 모델 – 37B activated for each token
- MoE(Mixture-of-Experts) architecture
- MLA(Multi-head Latent Attention) – KV cache 축소
- RoPE



DeepSeekMath - GRPO

- 수학 도메인에서 뛰어나게 동작할 수 있는 데이터 구축
- GRPO(Group Relative Policy Optimization) 도입



DeepSeekMath - GRPO

- GRPO(Group Relative Policy Optimization) 도입

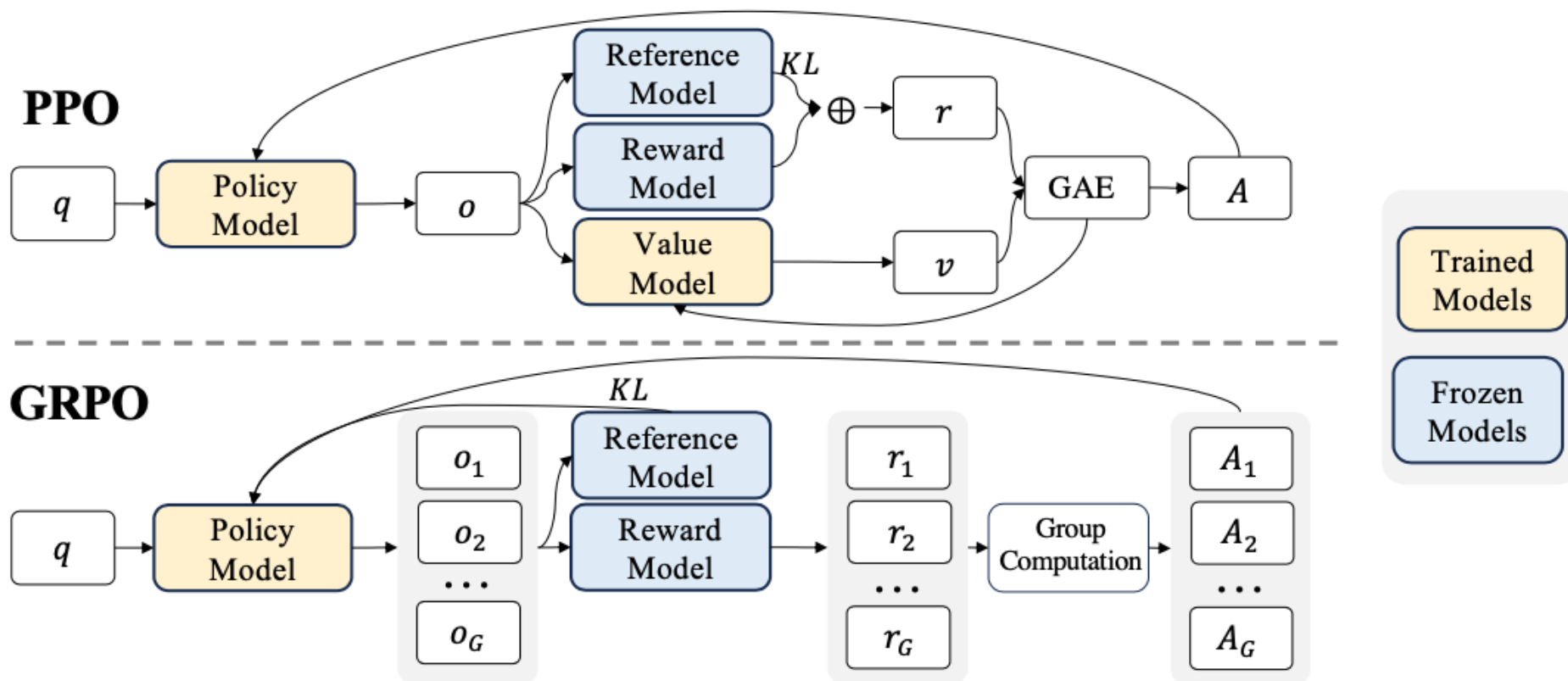
$$\mathcal{J}_{PPO}(\theta) = \mathbb{E}[q \sim P(Q), o \sim \pi_{\theta_{old}}(O|q)] \frac{1}{|o|} \sum_{t=1}^{|o|} \min \left[\frac{\pi_{\theta}(o_t|q, o_{<t})}{\pi_{\theta_{old}}(o_t|q, o_{<t})} A_t, \text{clip} \left(\frac{\pi_{\theta}(o_t|q, o_{<t})}{\pi_{\theta_{old}}(o_t|q, o_{<t})}, 1 - \epsilon,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

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

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

DeepSeekMath - GRPO

- GRPO(Group Relative Policy Optimization) 도입

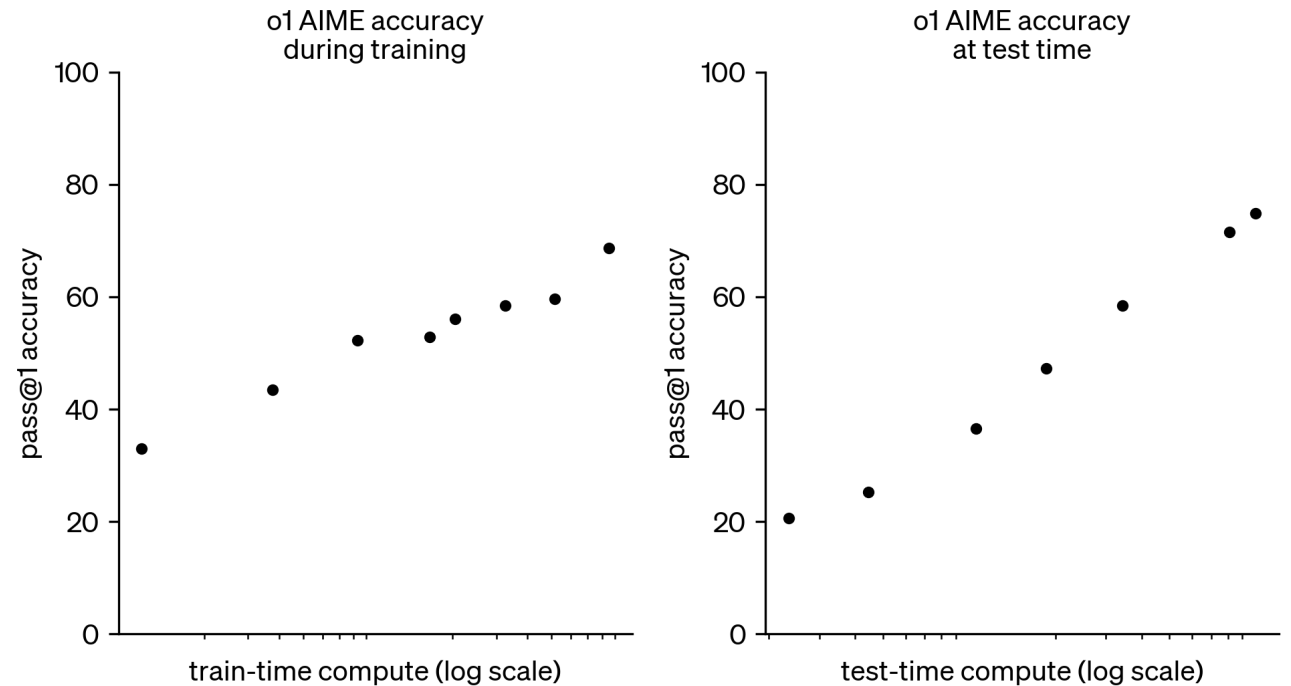


DeepSeek-R1-zero

- System prompt를 통한 CoT + GRPO
- CoT에서 생각하는 법을 스스로 배울 수 있도록 강화학습
- 정답이 있는 수학문제 등으로 강화학습 보상 설계
- 학습을 진행함에 따라 모델이 스스로 생각하는 모습이 관찰됨

Test-time scaling

- OpenAI의 blog posting - Learning to reason with LLMs
- 모델의 Inference 연산량이 많아짐에 따라 reasoning 성능이 증대됨을 서술
- 하지만, 방법은 모름
- CoT, ToT, GoT, BeamSearch 등 다양한 방식
- 최근 s1 방법도 제시됨



DeepSeek-R1-zero

- System prompt – CoT 기반 모델이 생각할 수 있는 기반 마련
- 사용시 응용, 강화학습에 유용하도록 제시

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.

DeepSeek-R1-zero: GRPO + reward

- 정확도 기반 보상 (Accuracy Reward): 수학 문제 정답 여부, 코드 실행 결과 등을 평가
- 형식 기반 보상 (Format Reward): 모델이 reasoning 프로세스를 <think> </think> 태그 내에서 작성하도록 유도하여 CoT를 수행, 정답 반환하도록 유도

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

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

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

DeepSeek-R1-zero

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

Table 2 | Comparison of DeepSeek-R1-Zero and OpenAI o1 models on reasoning-related benchmarks.

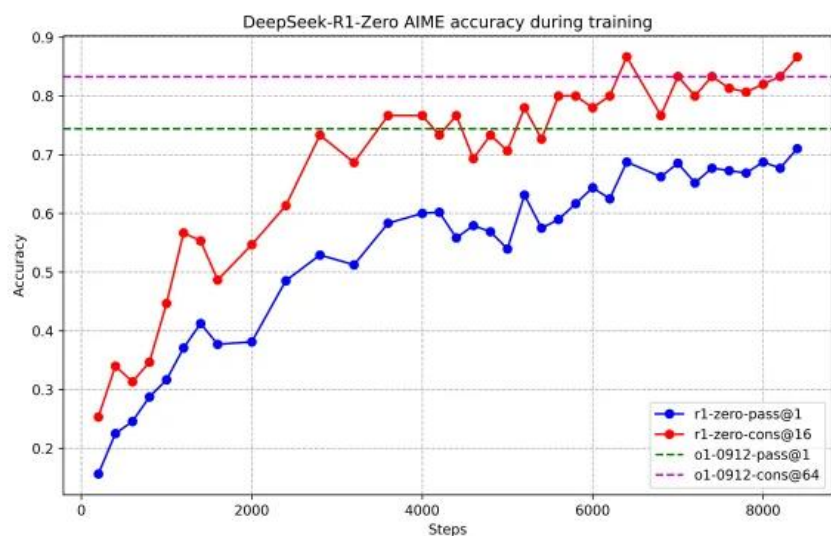
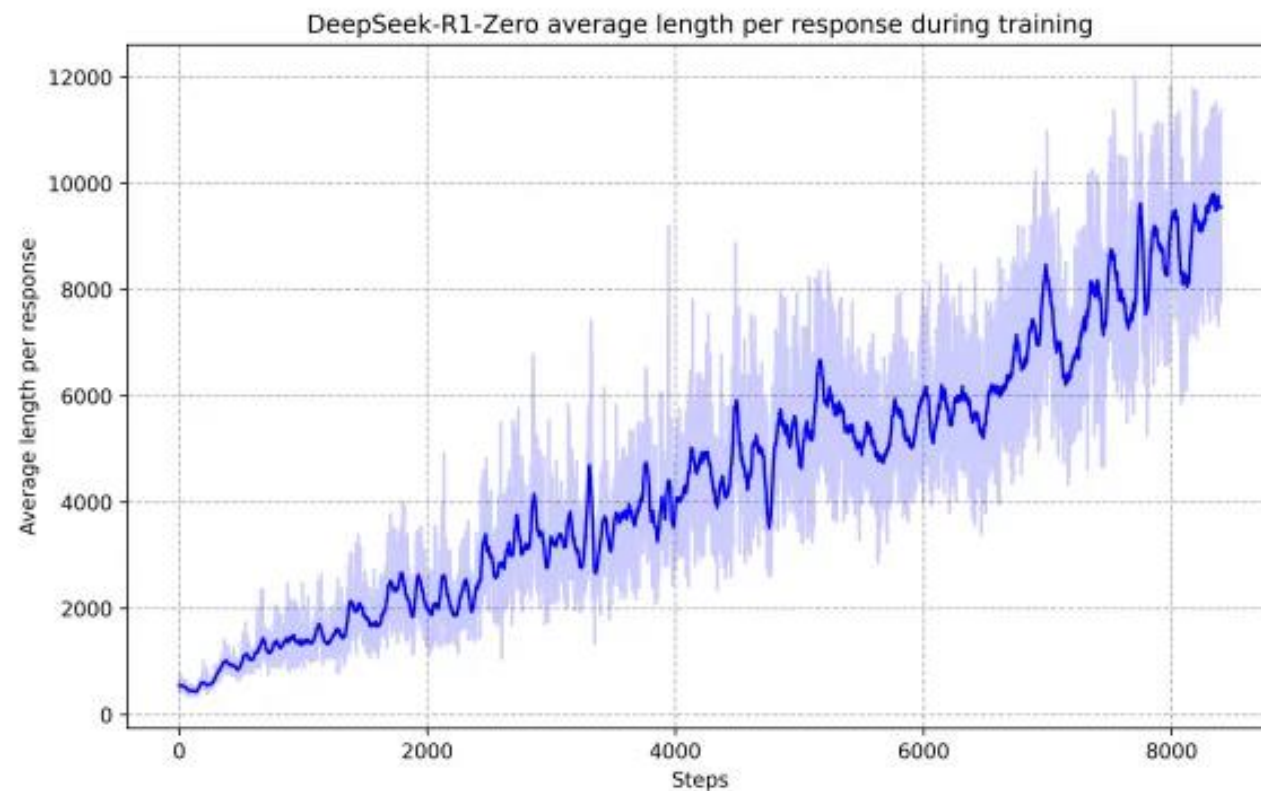


Figure 2 | AIME accuracy of DeepSeek-R1-Zero during training. For each question, we sample 16 responses and calculate the overall average accuracy to ensure a stable evaluation.



DeepSeek-R1-zero

Model	AIME 2024
	pass@1
OpenAI-o1-mini	63.6
OpenAI-o1-0912	74.4
DeepSeek-R1-Zero	71.0

Table 2 | Comparison of DeepSeek-R1-Zero with OpenAI o1 models

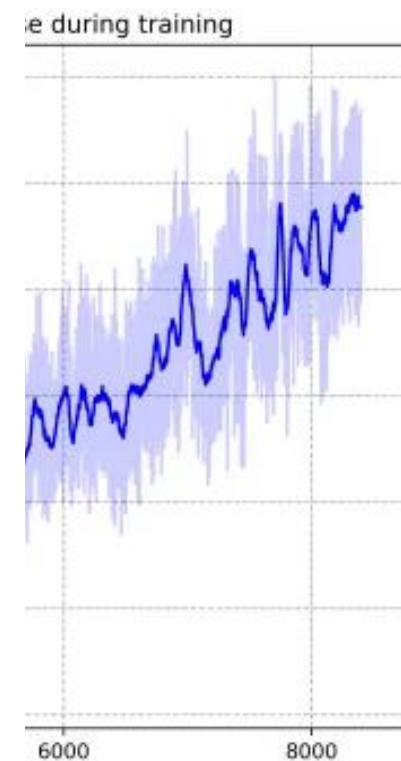
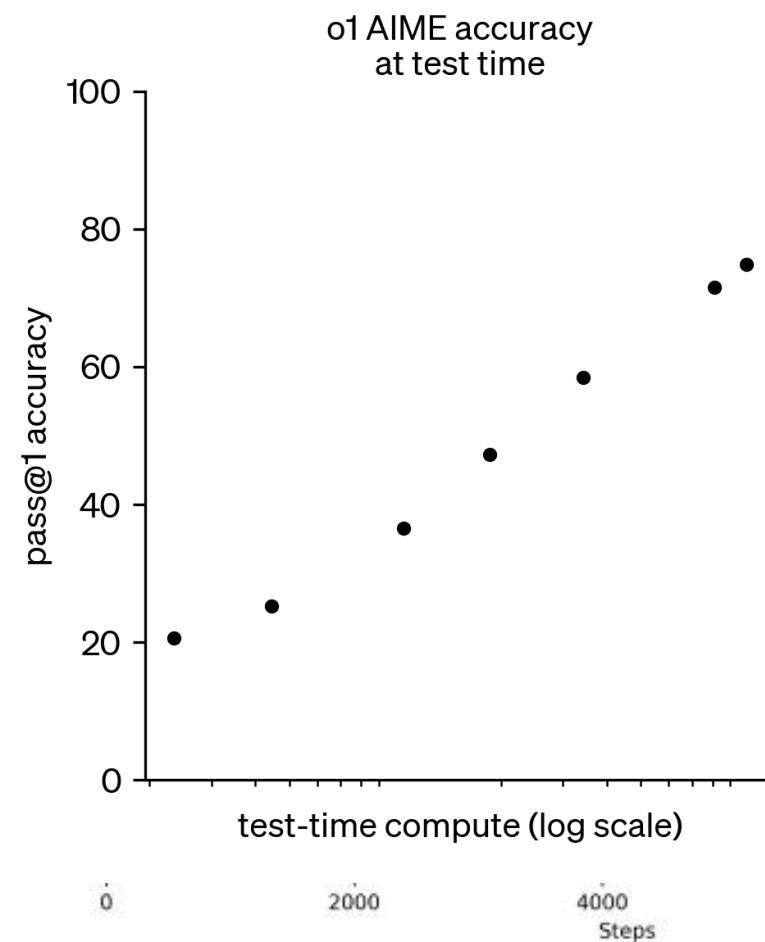
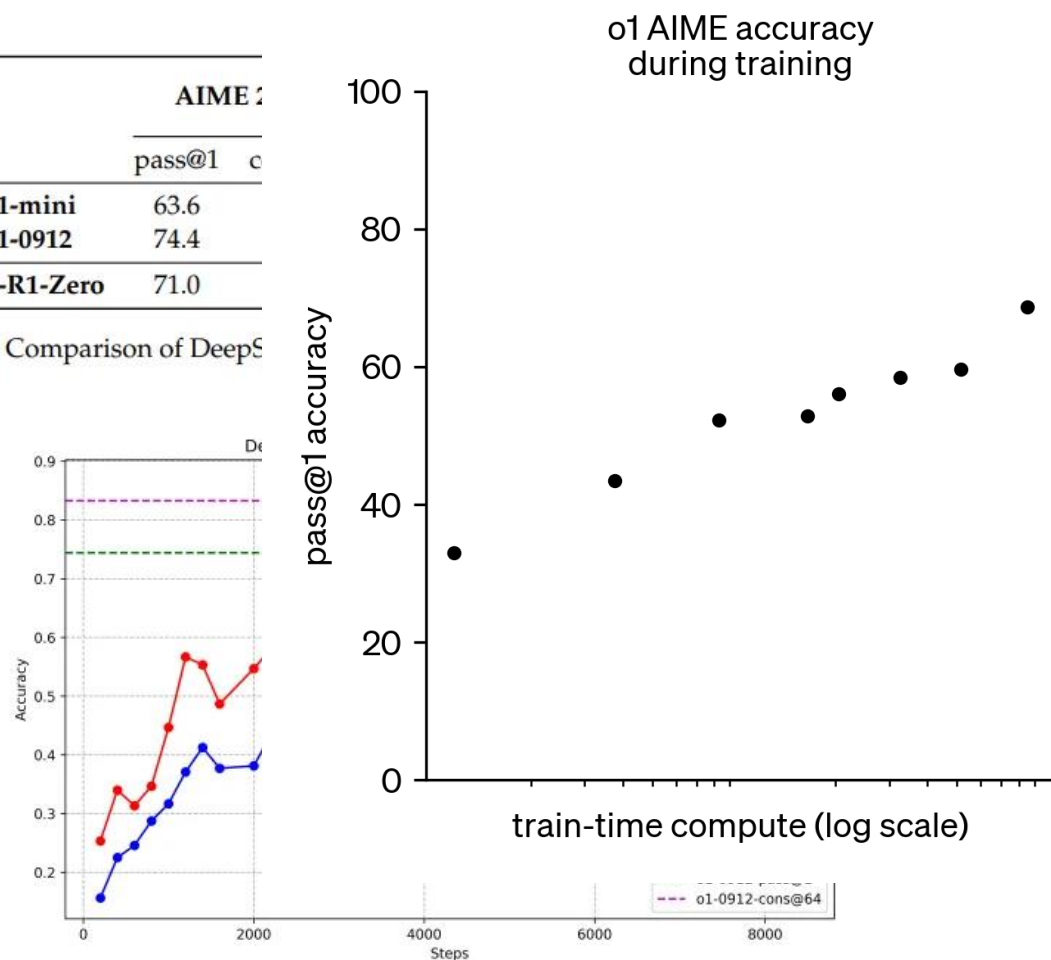


Figure 2 | AIME accuracy of DeepSeek-R1-Zero during training. For each question, we sample 16 responses and calculate the overall average accuracy to ensure a stable evaluation.

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 \dots

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

Rearrange to isolate the inner square root term:

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

\dots

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 \dots

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: \dots

\dots

Q & A

DeepSeek-R1

- 소량의 고품질 데이터를 초기 학습(cold start) 단계에서 활용함으로써 reasoning 성능을 더욱 향상시키거나 수렴 속도를 가속화할 수 있을까?
- 명확하고 일관된 Chain-of-Thought(CoT)를 생성할 뿐만 아니라, 강력한 범용 능력까지 갖춘 사용자 친화적인 모델을 어떻게 훈련할 수 있을까?

DeepSeek-R1

- Cold Start
- Reasoning-oriented Reinforcement Learning
- Rejection Sampling and Supervised Fine-Tuning
- Reinforcement Learning for all Scenarios

DeepSeek-R1: Cold Start

- CoT를 예제로 포함한 few-shot prompting 기법 사용하여
반성(reflection)과 검증(verification)이 포함된 상세한 답변 생성
- 위 데이터로 Fine-Tuning
- 결과
 - 가독성 향상: 출력 형식 - |special_token|<reasoning_process>|special_token|<summary>
 - 성능 일부 향상

DeepSeek-R1: Reasoning-oriented Reinforcement Learning

- R1-zero에서 진행된 강화학습 진행
- 보상에 언어 통일성 등의 특징을 더해서, 사용자 친화를 더함

DeepSeek-R1: Rejection Sampling and Supervised Fine-Tuning

- AGI / Multi-tasking을 위한 학습데이터 구축 및 학습
- Reasoning(60만) + Non-Reasoning(20만) 80만개의 데이터를
구축하여 V3-Base 모델에 2epochs 학습

DeepSeek-R1: Reinforcement Learning for all Scenarios

- 앞서 사용한 Pure RL에 Alignment 학습을 더하여, 사용자 친화적, 안전한 모델 설계
- Helpfulness, Harmlessness등을 학습

Q & A

Distillation

- DeepSeek-R1에서 생성한 80만 개의 샘플을 사용하여 오픈소스 모델을 직접 미세 조정(fine-tuning)
 - Qwen 시리즈: Qwen2.5-Math-1.5B, Qwen2.5-Math-7B, Qwen2.5-14B, Qwen2.5-32B
 - Llama 시리즈: Llama-3.1-8B, Llama-3.3-70B-Instruct

Distillation

- 단순한 지도학습(SFT) 기반의 증류(distillation) 기법만으로도 작은 모델의 reasoning 능력이 크게 향상
- 지식 증류 과정에서는 강화학습(RL) 단계를 적용하지 않았으며, 오직 지도학습(SFT)만 수행
- 강화학습을 활용한 소형 모델 성능 향상은 향후 연구 커뮤니티에서 추가로 탐색할 과제

Distillation

- 단순한 지도학습(SFT) 기반의 증류(distillation) 기법만으로도 작은 모델의 reasoning 능력이 크게 향상
- 지식 증류 과정에서는 강화학습(RL) 단계를 적용하지 않았으며, 오직 지도학습(SFT)만 수행
- 강화학습을 활용한 소형 모델 성능 향상은 향후 연구 커뮤니티에서 추가로 탐색할 과제

Models	HRM8K		
	GSM8K	MATH	OMNI_MATH
GPT-4o	91.2	74.5	30.7
GPT-4o-Mini	87.6	70.7	26.5
Qwen2.5-72B-Instruct	90.1	72.1	31.0
OLAFv2 (Reasoning Mode)	80.7	61.9	21.0
OLAFv2-Mini (Reasoning Mode)	68.1	50.8	17.6
Ko-R1-1.5B-Preview	43.3	73.1	29.8

Experiment

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

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

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

Table 6 | Comparison of distilled and RL Models on Reasoning-Related Benchmarks.

Q & A

감사합니다 😊