

2025-2 랩 세미나

Reinforcement Learning

이승윤

DAPO: An Open-Source LLM Reinforcement Learning System at Scale



¹ByteDance Seed ²Institute for AI Industry Research (AIR), Tsinghua University

³The University of Hong Kong

⁴SIA-Lab of Tsinghua AIR and ByteDance Seed

Full author list in Contributions

GRPO

GRPO는 가치 함수 없이 그룹 내 정규화된 보상 $\{R_i\}_{i=1}^G$ 으로 어드밴티지 $\hat{A}_{i,t}$ 를 추정한다.

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{(q,a) \sim \mathcal{D}, \{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \left(\min \left(r_{i,t}(\theta) \hat{A}_{i,t}, \text{clip} \left(r_{i,t}(\theta), 1 - \varepsilon, 1 + \varepsilon \right) \hat{A}_{i,t} \right) - \beta D_{\text{KL}}(\pi_{\theta} || \pi_{\text{ref}}) \right) \right]$$

$$\hat{A}_{i,t} = \frac{r_i - \text{mean}(\{R_i\}_{i=1}^G)}{\text{std}(\{R_i\}_{i=1}^G)}.$$

$$\text{clip} \left(\frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}, 1 - \varepsilon, 1 + \varepsilon \right)$$

$$r_{i,t}(\theta) = \frac{\pi_{\theta}(o_{i,t} | q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} | q, o_{i,<t})}.$$

A_i 는 i 번째 정책에 주어지는 advantage이며, 강화학습의 환경에서 주어지는 절대적인 보상 reward와는 조금 다른 개념, 실제로는 상대적인 보상, 이득을 의미

clip 함수는 원하는 값이 너무 크거나 작아지지 않도록 특정 범위 안으로 가두어 버리는 역할
=> 학습 안정화

현재 정책(π_{θ})과 참조 정책($\pi_{\theta_{\text{old}}}$)의 비율 r_i

=> 이 비율은 지난번 정책에 비했을 때, 새로운 정책이 선택될 가능성

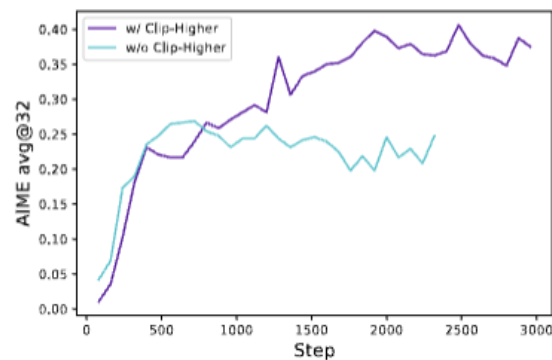
- 정책이 업데이트될 때 특정 행동을 선택할 확률이 얼마나 달라졌는지(증가/감소 비율)를 직접 반영할 수 있게됨
- rule-based로 계산하는 직접 보상(Reward)을 그룹 단위로 표준화한 상대 보상(Advantage)을 사용하여 비용 효율적인 강화학습을 실현
- **비율(r)과 Advantage(A)를 곱해주는 이유: 정책이 바뀌었을 때 그 바뀐 정도가 실제로 그 행동의 이득(Advantage)에 부합하는지를 평가하기 위함**
- clip을 통해 비율 자체를 특정 범위 안으로 가두어 버리면, A_i 를 곱하면서 너무 크게 보상하거나 과하게 불이익을 주게 되는 상황을 방지할 수 있고, 현재 정책이 이전 정책에서 과도하게 멀어지지 않게 됨

Problems

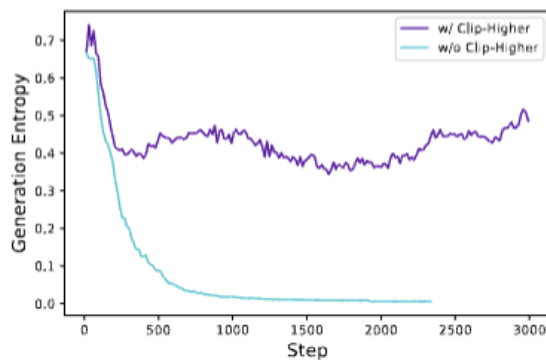
- GRPO baseline suffers from several key issues such as entropy collapse, reward noise, and training instability
- **4 Strategy**
 - 1. Clip-Higher, which promotes the diversity of the system and avoids entropy collapse**
 - 2. Dynamic Sampling, which improves training efficiency and stability**
 - 3. Token-Level Policy Gradient Loss, which is critical in long-CoT RL scenarios**
 - 4. Overlong Reward Shaping, which reduces reward noise and stabilizes training**

Clip-Higher

- GRPO를 사용한 초기 실험에서 정책의 엔트로피가 학습이 진행됨에 따라 빠르게 감소하는 엔트로피 붕괴 현상을 관찰
 - 특정 그룹의 샘플링된 응답은 거의 동일한 경향 => 제한된 탐색과 초기 결정적 정책의 오류로 인한 붕괴
- ⇒ ϵ_{high} 을 높여 (논문에서는 0.28 사용) 낮은 확률 토큰의 확률 증가에 더 많은 여유를 주어 정책 엔트로피를 높이고 다양한 샘플 생성을 촉진
- ⇒ 낮은 확률의 토큰 증가에 더 많은 여지



(a) Accuracies on AIME.



(b) Entropy of actor model.

Figure 2 The accuracy on the AIME test set and the entropy of the actor model's generated probabilities during the RL training process, both before and after applying **Clip-Higher** strategy.

$$\mathcal{J}_{\text{DAPO}}(\theta) = \mathbb{E}_{(q,a) \sim \mathcal{D}, \{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{\sum_{i=1}^G |o_i|} \sum_{i=1}^G \sum_{t=1}^{|o_i|} \min \left(r_{i,t}(\theta) \hat{A}_{i,t}, \text{clip} \left(r_{i,t}(\theta), 1 - \epsilon_{\text{low}}, 1 + \epsilon_{\text{high}} \right) \hat{A}_{i,t} \right) \right]$$

Dynamic Sampling

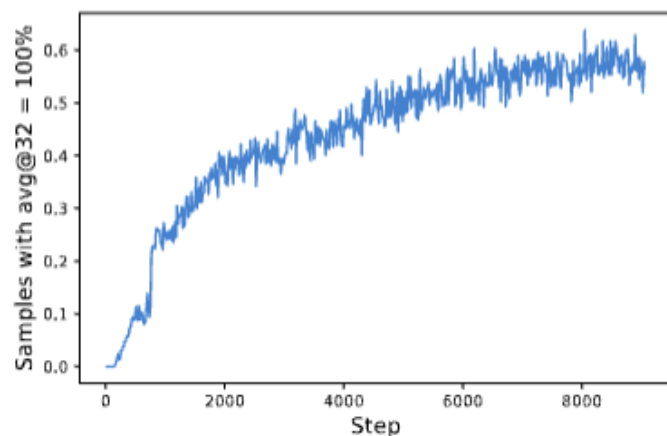
- 일부 프롬프트가 정확도 1 (Figure 3b) 또는 0을 달성하여 그룹 어드밴티지가 0

⇒ 효과적인 기울기 신호가 약해지고 배치(batch)의 효율성이 저하

- 특정 그룹의 샘플링된 응답은 거의 동일한 경향 => 제한된 탐색과 초기 결정적 정책의 오류로 인한 붕괴

⇒ 정확도가 1과 0인 프롬프트를 과도하게 샘플링하고 필터링하여 배치의 모든 프롬프트에 유효한 기울기를 남기고 일관된 수의 프롬프트를 유지할 것을 제안

⇒ 훈련 전에 정확도가 0 또는 1이 아닌 샘플로 배치가 완전히 채워질 때까지 계속 샘플링



(b) The proportion of samples with an accuracy of 1.

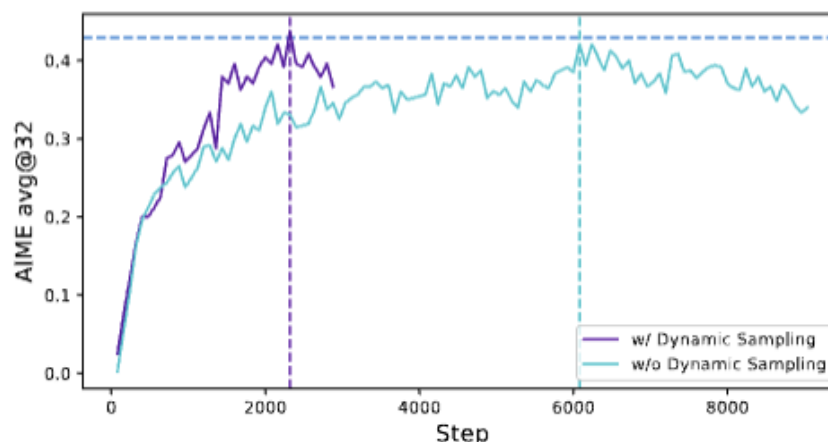
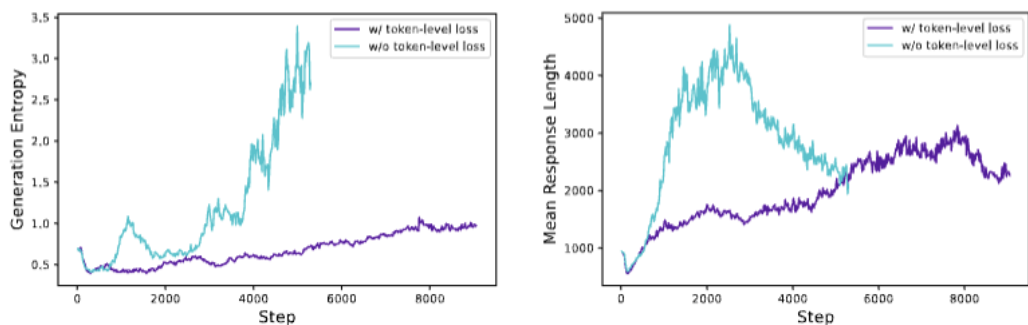


Figure 6 The training progress before and after applying dynamic sampling on a baseline setting.

Token-Level Policy Gradient Loss

- GRPO의 sample-level loss 계산 방식이 긴 응답 내 토큰의 기여도를 낮추고 과도하게 긴 저품질 응답을 효과적으로 제어하지 못하는 문제를 해결
- ⇒ 각 토큰이 전체 손실 계산에 동등하게 기여하도록 토큰 레벨 평균 (sum over tokens, then average over samples)을 사용
- ⇒ 이를 통해 긴 시퀀스가 전체 기울기 업데이트에 더 많은 영향을 미치고, 특정 생성 패턴이 응답 길이에 상관없이 동등하게 촉진되거나 억제
- ⇒ DAPO는 그룹 내 모든 시퀀스의 모든 토큰에 대한 손실을 합산한 다음, 이를 전체 토큰 수로 나누어 평균



(a) Entropy of actor model's generation probabilities. (b) Average length of actor model-generated responses
Figure 4 The entropy of the probability distribution of the actor model, as well as the changes in response length.

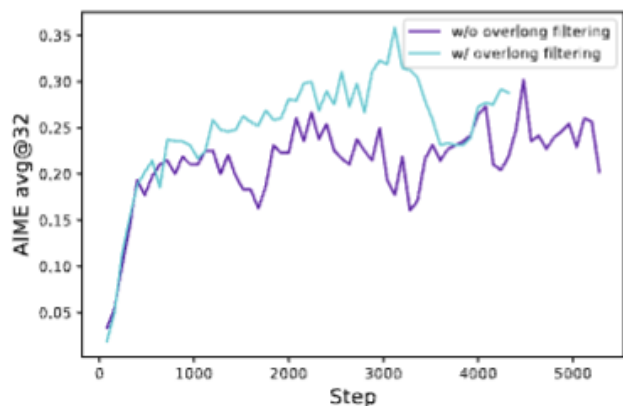
특징	GRPO의 Sample-Level Loss	DAPO의 Token-Level Loss
평균화 단위	시퀀스(샘플) 단위로 먼저 평균하고, 그 다음 그룹 내 시퀀스들을 평균	토큰 단위로 모든 토큰의 손실을 합산한 후 전체 토큰 수로 평균
시퀀스 길이의 영향	긴 시퀀스의 개별 토큰 기여도 낮음 (시퀀스가 동등한 가중치)	긴 시퀀스가 전체 그라디언트에 더 큰 영향 (토큰이 동등한 가중치)
목표	각 샘플에 대한 정책 업데이트의 균형 유지	각 토큰에 대한 정책 업데이트의 균형 유지, 긴 시퀀스의 중요성 강조
Long-CoT에서의 문제점	긴 추론 패턴 학습 저해, 저품질 긴 시퀀스에 대한 약한 페널티	
Long-CoT에서의 장점		긴 추론 패턴 학습 강화, 저품질 긴 시퀀스 효과적 페널티

Overlong Reward Shaping

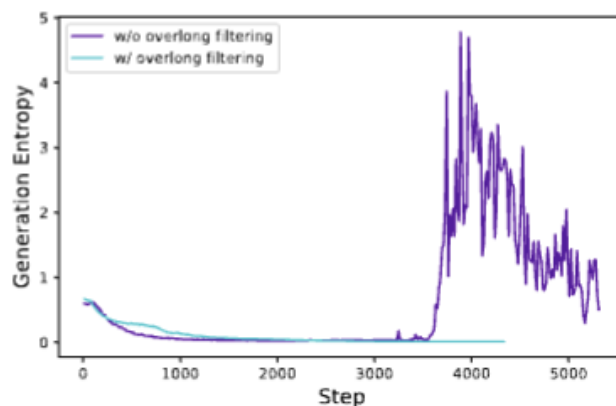
- RL 훈련에서 일반적으로 생성에 대한 최대 길이를 설정하고 과도하게 긴 샘플은 그에 따라 잘림
- 잘린 샘플에 대한 부적절한 보상 형성은 보상 노이즈를 유발하고 훈련 과정을 크게 방해할 수 있다는 것을 발견

⇒ 먼저 잘린 샘플의 손실을 마스킹하는 **Overlong Filtering** 전략을 적용

⇒ 특히, 응답 길이가 미리 정의된 최대값을 초과하면 처벌 간격을 정의. 이 간격 내에서 응답이 길수록 더 큰 처벌



(a) Performance on AIME.



(b) Entropy of actor model.

$$R_{\text{length}}(y) = \begin{cases} 0, & |y| \leq L_{\text{max}} - L_{\text{cache}} \\ \frac{(L_{\text{max}} - L_{\text{cache}}) - |y|}{L_{\text{cache}}}, & L_{\text{max}} - L_{\text{cache}} < |y| \leq L_{\text{max}} \\ -1, & L_{\text{max}} < |y| \end{cases}$$

Figure 5 The accuracy of the actor model on AIME and the entropy of its generation probabilities, both before and after applying **Overlong Reward Shaping** strategy.

Main Result

Table 1 Main results of progressive techniques applied to **DAPO**

Model	AIME24 _{avg@32}
DeepSeek-R1-Zero-Qwen-32B	47
Naive GRPO	30
+ Overlong Filtering	36
+ Clip-Higher	38
+ Soft Overlong Punishment	41
+ Token-level Loss	42
+ Dynamic Sampling (DAPO)	50

What to observe in RL?

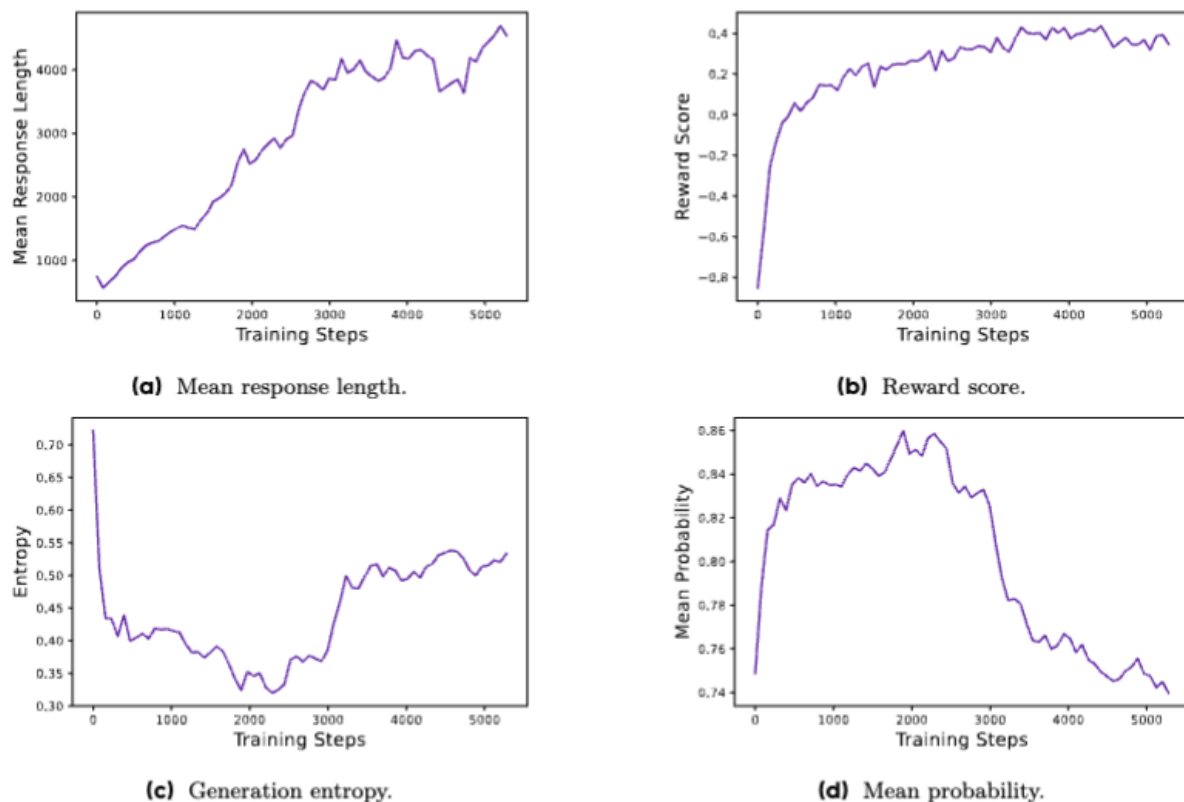


Figure 7 The metric curves of response length, reward score, generation entropy, and the mean probability of DAPO, which show the dynamics of RL training and serve as essential monitoring indicators to identify potential issues.

- 길이 증가는 모델에게 더 넓은 탐색 공간을 제공하여 더 복잡한 추론 행동을 샘플링하고 훈련을 통해 점진적으로 강화할 수 있도록 함

- 보상 증가 추세는 비교적 안정적이며 실험 설정 조정으로 인해 크게 변동하거나 감소하지 않음

- 모델의 엔트로피는 적절한 범위 내에서 유지되어야 함
 - 엔트로피가 지나치게 낮으면 확률 분포가 지나치게 날카로워져 탐색 능력을 잃음
 - 반대로 엔트로피가 지나치게 높으면 횡설수설 및 반복 생성과 같은 과도한 탐색 문제가 발생

S²R: Teaching LLMs to Self-verify and Self-correct via Reinforcement Learning



**Ruotian Ma^{1*}, Peisong Wang^{2*}, Cheng Liu¹, Xingyan Liu¹,
Jiaqi Chen³, Bang Zhang¹, Xin Zhou⁴, Nan Du^{1†}, Jia Li^{5†}**

¹Tencent ²Tsinghua University

³The University of Hong Kong ⁴Fudan University

⁵The Hong Kong University of Science and Technology (Guangzhou)

ruotianma@tencent.com, wps22@mails.tsinghua.edu.cn

S2R

- 추론 과정에서 모델이 자체 검증하고 자체 수정하도록 학습시켜 LLM 추론을 향상시키는 효율적인 프레임워크
 - 먼저 신중하게 선별된 데이터에 대한 지도 학습을 통해 반복적인 자체 검증 및 자체 수정 행동으로 LLM을 초기화
 - 결과 수준 및 프로세스 수준 강화 학습을 통해 자체 검증 및 자체 수정 기술을 더욱 강화
- ⇒ 스스로 생성한 틀린 오답에 대해 스스로 수정할 수 있도록 유도
- ⇒ S2R은 자기 검증 및 자기 수정이라는 두 가지 중요한 사고 기술을 반복적으로 채택하여 LLM이 깊이 생각하도록 가르치는 데 중점

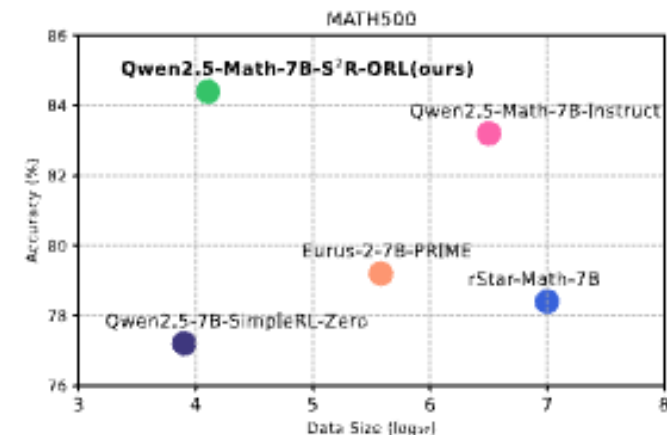
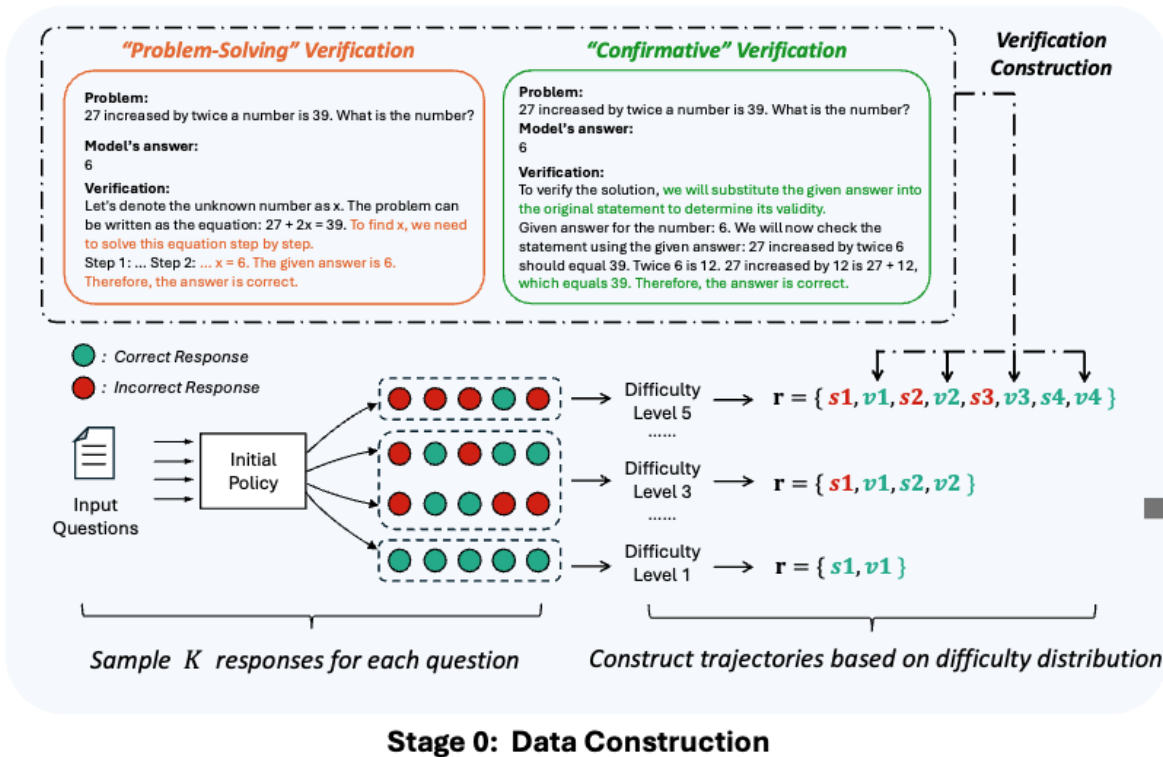


Figure 1: The data efficiency of S²R compared to competitive baseline methods.

Step0: Data Construction



Solve, Verify, End

- Input question에 대해 K번의 응답 생성을 통해 정답여부를 판단
⇒ 많이 맞춘 input은 난이도가 쉬운, 반대는 어려운 것으로 간주
- 난이도에 따라 추론 궤적을 다양화
⇒ 쉬운 것은 바로 $s1, v1$, end 되도록 / 어려운 것은 실패와 검증을 반복하도록 구성
- 예시: $s1, v1, s2, v2, s3, v3$, end인 경우
⇒ $s1, s2$ 는 모델이 생성한 틀린 answer
⇒ $v1, v2$ 는 GPT-4의 verification / $s3$ 는 고품질 정답 텍스트

Step0: Data Construction

LLM의 목표 출력 (모델이 모방해야 할 행동 시퀀스 y):

[s1] (첫 번째 **solve** 행동 - 오답)

"이 문제를 풀어보겠습니다. 처음에 사과는 10개 있었습니다. 당신이 3개를 먹었으므로, $10 + 3 = 13$ 개입니다. 친구가 5개를 더 주었으니, $13 + 5 = 18$ 개입니다. 따라서 사과의 총 개수는 **18**개입니다."

[v1] (첫 번째 **verify** 행동 - s_1 이 오답임을 정확히 판단)

"잠깐, 제 풀이를 다시 확인해 보겠습니다."

검증: 제시된 답은 18개입니다. 문제에서 '3개를 먹었다'는 것은 사과를 빼는 행동을 의미합니다. 만약 10개에서 3개를 먹었다면 $10 - 3 = 7$ 개가 됩니다. 그 후에 친구가 5개를 더 주었다면 $7 + 5 = 12$ 개가 되어야 합니다. 18이라는 계산은 3개를 먹은 것을 잘못 더하기로 계산한 것입니다.

따라서, 이 답은 **incorrect**입니다. 다시 시도해 보겠습니다."

[s2] (두 번째 **solve** 행동 - 오답, 첫 번째 오류는 수정했으나 새로운 오류 발생)

"다시 풀어보겠습니다. 처음에 사과는 10개 있었습니다. 당신이 3개를 먹었으니, $10 - 3 = 7$ 개입니다. 친구가 5개를 더 주었으므로, $7 \times 5 = 35$ 개입니다. 따라서 사과의 총 개수는 **35**개입니다."

[v2] (두 번째 **verify** 행동 - s_2 가 오답임을 정확히 판단)

"잠깐, 제 풀이를 다시 확인해 보겠습니다."

검증: 제시된 답은 35개입니다. 이전 단계에서 $10 - 3 = 7$ 까지는 옳았습니다. 그러나 '친구가 5개를 더 주었다'는 것은 5를 더하는 것을 의미하지 곱하는 것을 의미하지 않습니다. $7 + 5 = 12$ 가 되어야 합니다.

따라서, 이 답은 **incorrect**입니다. 다시 시도해 보겠습니다."

[s3] (세 번째 **solve** 행동 - 정답)

"세 번째 시도입니다. 처음에 사과는 10개 있었습니다. 당신이 3개를 먹었으므로, $10 - 3 = 7$ 개입니다. 친구가 5개를 더 주었으니, $7 + 5 = 12$ 개입니다. 따라서 사과의 총 개수는 **12**개입니다."

[v3] (세 번째 **verify** 행동 - s_3 이 정답임을 정확히 판단)

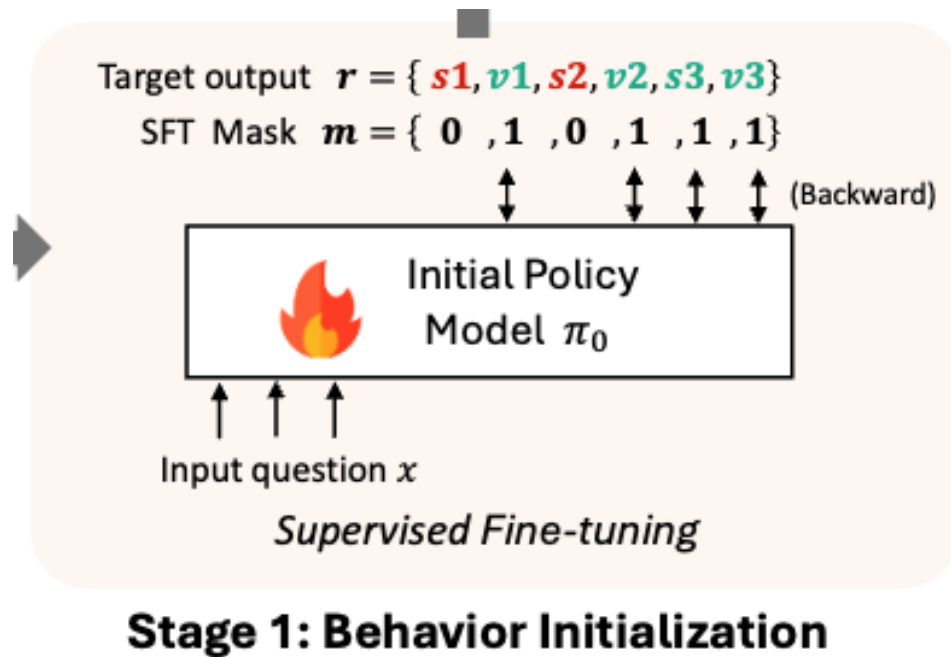
"잠깐, 제 풀이를 다시 확인해 보겠습니다."

검증: 제시된 답은 12개입니다. 처음에 10개에서 3개를 먹었으므로 $10 - 3 = 7$ 개가 남습니다. 친구가 5개를 더 주었으므로 $7 + 5 = 12$ 개가 됩니다. 이 계산은 문제의 조건에 정확히 부합하며, 논리적으로 타당합니다.

따라서, 이 답은 **correct**입니다."

[<end>] (추론 종료)

Step1: SFT



일종의 cold start: Masked SFT

- 모델이 초기에 틀리게 생성했던 부분은 loss 계산하지 않음
- GPT-4가 verification한 부분과 행동 수정을 통해 올바르게 정답을 맞춘 부분만 loss를 계산해서 SFT

⇒ Verification과 self correction의 반복을 배우는 효과

- 해당 모델은 Behavior Init 모델이라고 명명

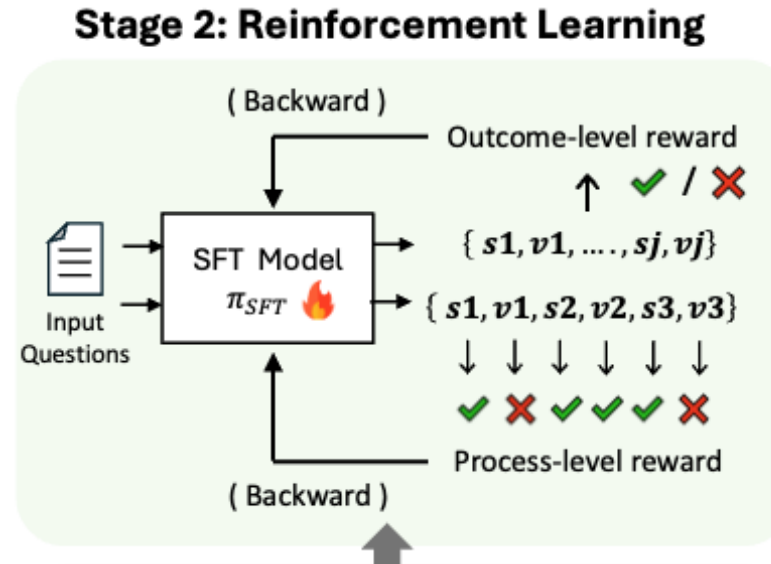
Step2: Two-way RL

$$\hat{b}(x, y^{(m)}) = \frac{1}{M-1} \sum_{j=1, j \neq m}^M R(x, y^{(j)})$$

Outcome-level RLOO

GRPO와 유사하지만, 해당 rollout 외의 rollout들의 보상을 평균내어 비교하는 방식으로 advantage를 계산

- 최종 solution이 정답을 맞추면 1, 아니면 -1로 보상



Process-level GRPO

- solve 행동에 대한 보상:

생성된 solution이 실제 정답과 일치하면 1점, 틀리면 -1점

- verify 행동에 대한 보상:

모델의 자체 검증이 직전의 solve 행동의 실제 정답 여부를 정확하게 판단하면 1점, 틀리면 -1점

Main Experiments

Model	MATH 500	AIME 2024	AMC 2023	College Math	Olympiad Bench	GSM8K	GaokaoEn 2023	Average
<i>Frontier LLMs</i>								
GPT-4o	76.6	9.3	47.5	48.5	43.3	92.9	67.5	55.1
GPT-o1-preview	85.5	44.6	90.0	-	-	-	-	-
GPT-o1-mini	90.0	56.7	95.0	57.8	65.3	94.8	78.4	76.9
<i>Top-tier Open-source Reasoning LLMs</i>								
NuminaMath-72B-CoT	64.0	3.3	70.0	39.7	32.6	90.8	58.4	51.3
LLaMA3.1-70B-Instruct	65.4	23.3	50.0	42.5	27.7	94.1	54.0	51.0
Qwen2.5-Math-72B-Instruct	85.6	30.0	70.0	49.5	49.0	95.9	71.9	64.6
<i>General Model: Llama-3.1-8B-Instruct</i>								
Llama-3.1-8B-Instruct	48.0	<u>6.7</u>	<u>30.0</u>	30.8	15.6	84.4	41.0	36.6
Llama-3.1-8B-Instruct + Original Solution SFT	31.0	3.3	7.5	22.0	8.0	58.7	28.3	22.7
Llama-3.1-8B-Instruct + Long CoT SFT	51.4	<u>6.7</u>	27.5	36.3	<u>19.0</u>	<u>87.0</u>	48.3	<u>39.5</u>
Llama-3.1-8B-S²R-BI (ours)	49.6	10.0	20.0	33.3	17.6	85.3	41.0	36.7
Llama-3.1-8B-S²R-PRL (ours)	<u>53.6</u>	<u>6.7</u>	25.0	<u>33.7</u>	18.5	86.7	43.1	38.2
Llama-3.1-8B-S²R-ORL (ours)	55.0	<u>6.7</u>	32.5	34.7	20.7	87.3	<u>45.2</u>	40.3
<i>General Model: Qwen2-7B-Instruct</i>								
Qwen2-7B-Instruct	51.2	3.3	30.0	18.2	19.1	86.4	39.0	35.3
Qwen2-7B-Instruct + Original Solution SFT	41.2	0.0	25.0	30.1	10.2	74.5	34.8	30.8
Qwen2-7B-Instruct + Long CoT SFT	60.4	<u>6.7</u>	32.5	36.3	23.4	81.2	53.5	42.0
Qwen2-7B-S²R-BI (ours)	61.2	3.3	27.5	41.1	27.1	<u>87.4</u>	49.1	42.4
Qwen2-7B-S²R-PRL (ours)	65.4	<u>6.7</u>	<u>35.0</u>	<u>36.7</u>	<u>27.0</u>	89.0	<u>49.9</u>	44.2
Qwen2-7B-S²R-ORL (ours)	<u>64.8</u>	3.3	42.5	34.7	26.2	86.4	50.9	<u>44.1</u>
<i>Math-Specialized Model: Qwen2.5-Math-7B</i>								
Qwen2.5-Math-7B	51.0	16.7	45.0	21.5	16.7	58.3	39.7	35.6
Qwen2.5-Math-7B-Instruct	83.2	13.3	72.5	47.0	40.4	95.6	67.5	59.9
Eurus-2-7B-PRIME (Cui et al., 2025)	79.2	<u>26.7</u>	57.8	45.0	42.1	88.0	57.1	56.6
rStar-Math-7B ² (Guan et al., 2025)	78.4	<u>26.7</u>	47.5	52.5	47.1	89.7	65.7	58.2
Qwen2.5-7B-SimpleRL (Zeng et al., 2025)	82.4	<u>26.7</u>	62.5	-	43.3	-	-	-
Qwen2.5-Math-7B + Original Solution SFT	58.0	6.7	42.5	35.8	20.0	79.5	51.9	42.1
Qwen2.5-Math-7B + Long CoT SFT	80.2	16.7	60.0	<u>49.6</u>	42.1	91.4	69.1	58.4
Qwen2.5-Math-7B-S²R-BI (ours)	81.6	23.3	60.0	43.9	44.4	91.9	<u>70.1</u>	59.3
Qwen2.5-Math-7B-S²R-PRL (ours)	<u>83.4</u>	<u>26.7</u>	<u>70.0</u>	43.8	<u>46.4</u>	<u>93.2</u>	70.4	<u>62.0</u>
Qwen2.5-Math-7B-S²R-ORL (ours)	84.4	23.3	77.5	43.8	44.9	92.9	<u>70.1</u>	62.4

Table 2: The performance of S²R and other strong baselines on the most challenging math benchmarks is presented. **BI** refers to the behavior-initialized models through supervised fine-tuning, **ORL** denotes models trained with outcome-level RL, and **PRL** refers to models trained with process-level RL. The highest results are highlighted in **bold** and the second-best results are marked with underline.

Stage 1: Behavior Initialization		
Base Model	Source	# Training Data
Llama-3.1-8B-Instruct	MATH	4614
Qwen2-7B-Instruct	MATH	4366
Qwen2.5-Math-7B	MATH	3111
Stage 2: Reinforcement Learning		
Base Model	Source	# Training Data
Llama-3.1-8B-Instruct	MATH+GSM8K	9601
Qwen2-7B-Instruct	MATH+GSM8K	9601
Qwen2.5-Math-7B	MATH+OpenMath2.0	10000

Table 1: Training data statistics.

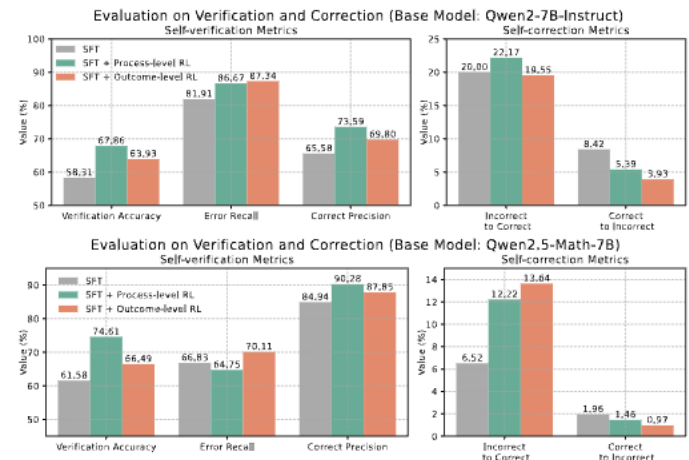


Figure 3: Evaluation on verification and correction.

- Stage 1의 SFT만으로도 좋은 성능
- ⇒ QwQ에서 추출한 Long CoT를 활용한 SFT보다 효과적
- 모델이 틀린 답변을 수정하는 능력이 RL 훈련 후 크게 향상
- ⇒ 모델이 올바른 답변을 실수로 변경하는 비율도 현저히 감소
- Process Level보다 **Outcome Level RL**이 더 효과적임
- ⇒ 결과 수준의 RL이 중간 정확성을 강조하지 않고 탐색할 수 있도록 하여 장

Self Correction Arise in Different Difficulty Problem

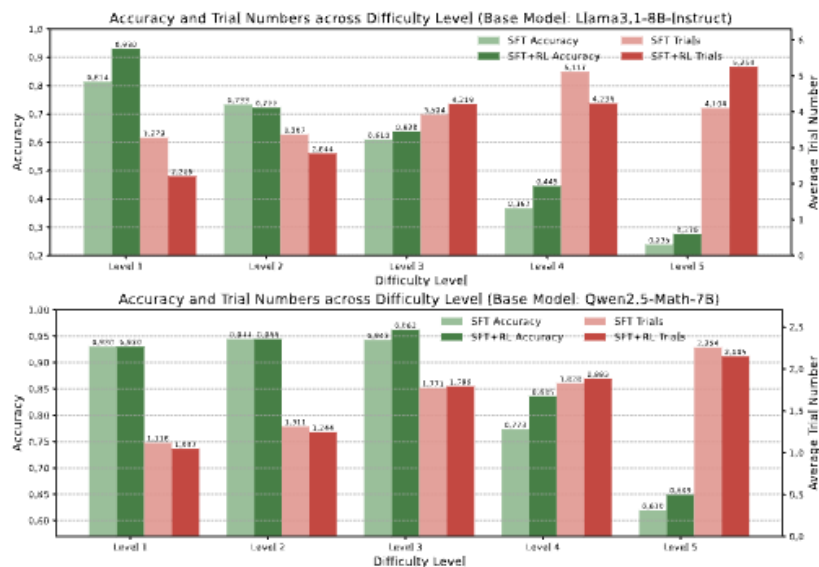


Figure 4: The accuracy and average trial number of different models across difficulty levels on MATH500.

- 모든 난이도 레벨에서 SFT+RL 모델(어두운 녹색)이 SFT 모델(밝은 녹색)보다 높은 정확도
 - 특히, 난이도가 높아질수록 SFT 모델의 정확도가 급격히 떨어지지만, SFT+RL은 이를 더 효과적으로 방어하며 어려운 문제에서도 성능을 개선
 - 난이도가 높은 문제(Level 3, 4, 5)에서는 SFT+RL이 SFT보다 더 많은 시도를 함으로써 정확도를 개선
- ⇒ 이는 RL이 모델이 더 어려운 문제에 대해 더 깊게 생각하고 필요한 경우 여러 번의 자기 검증 및 수정을 시도하도록 장려했음을 의미

모델은 난이도에 따라 추론 노력을 동적으로 할당하는 방법을 배우고, RL을 통해 더 어려운 문제에 대해 더 많은 자기 검증 및 수정을 시도

Expansion to Offline RL

Model	Datasets							Average
	MATH 500	AIME 2024	AMC 2023	College Math	Olympiad Bench	GSM8K	GaokaoEn 2023	
Qwen2-7B-Instruct	51.2	3.3	30.0	18.2	19.1	86.4	39.0	35.3
Qwen2-7B-S²R-BI (ours)	61.2	3.3	27.5	41.1	27.1	87.4	49.1	42.4
Qwen2-7B-S²R-PRL (ours)	65.4	<u>6.7</u>	35.0	36.7	<u>27.0</u>	89.0	<u>49.9</u>	<u>44.2</u>
Qwen2-7B-S²R-ORL (ours)	<u>64.8</u>	3.3	42.5	34.7	26.2	86.4	50.9	44.1
Qwen2-7B-Instruct-S²R-PRL-offline (ours)	61.6	10.0	32.5	40.2	26.5	<u>87.6</u>	50.4	44.1
Qwen2-7B-Instruct-S²R-ORL-offline (ours)	61.0	<u>6.7</u>	<u>37.5</u>	<u>40.5</u>	27.3	87.4	49.6	44.3
Qwen2.5-Math-7B	51.0	16.7	45.0	21.5	16.7	58.3	39.7	35.6
Qwen2.5-Math-7B-S²R-BI (ours)	81.6	<u>23.3</u>	60.0	43.9	44.4	91.9	70.1	59.3
Qwen2.5-Math-7B-S²R-PRL (ours)	<u>83.4</u>	26.7	<u>70.0</u>	43.8	<u>46.4</u>	93.2	<u>70.4</u>	<u>62.0</u>
Qwen2.5-Math-7B-S²R-ORL (ours)	84.4	23.3	77.5	43.8	44.9	<u>92.9</u>	70.1	62.4
Qwen2.5-Math-7B-S²R-PRL-offline (ours)	<u>83.4</u>	<u>23.3</u>	62.5	50.0	46.7	<u>92.9</u>	72.2	61.6
Qwen2.5-Math-7B-S²R-ORL-offline (ours)	82.0	20.0	67.5	<u>49.8</u>	45.8	92.6	<u>70.4</u>	61.2

Table 5: Comparison of S²R using online and offline RL training.

- online RL과 달리 offline RL에서는 process-level 이 outcome-level 보다 성능이 뛰어남
- ⇒ 모델이 동적 궤적을 자유롭게 탐색할 수 있도록 하는 데 탁월한 outcome-level RL은 online 파라미터 업데이트 중에 즉석에서 샘플링하는데 더 적합
- ⇒ 중간 단계에 대한 정확한 baseline 추정이 필요한 process-level RL은 더 큰 규모의 데이터 샘플링으로 더 정확한 baseline 추정치를 제공할 수 있는 offline 궤적 샘플링의 이점
- Offline RL은 대부분의 벤치마크에서 성능을 일관되게 향상시키고 online RL과 비교 가능한 결과를 달성