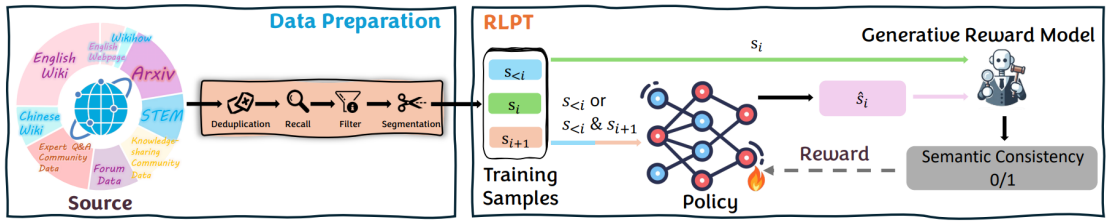
Reinforcement Learning on Pre-Training Data

250930

가짜연구소 허의주

Reinforcement Learning on Pre-Training Data

인간의 주석 없이 사전 훈련 데이터로부터 보상 신호를 도출하는 Next-Segment Reasoning를 통해, LLM의 일반화 가능 추론 능력을 강화하는 새로운 훈련 시간 확장 패러다임



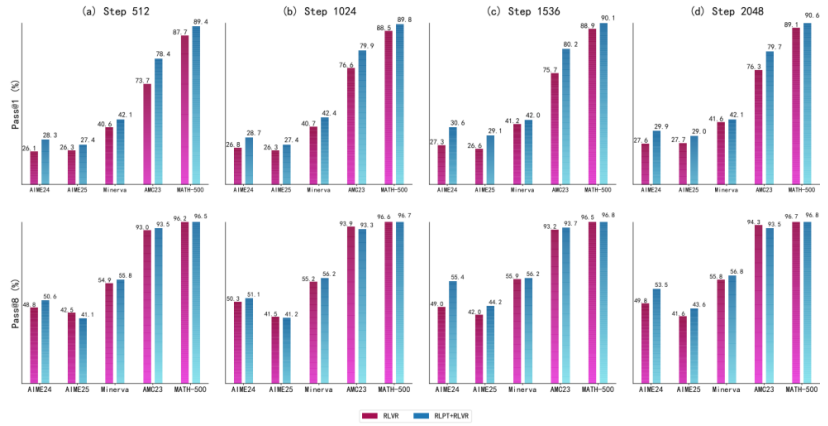


Figure 3: Comparative scaling properties of RLVR and RLPT + RLVR.

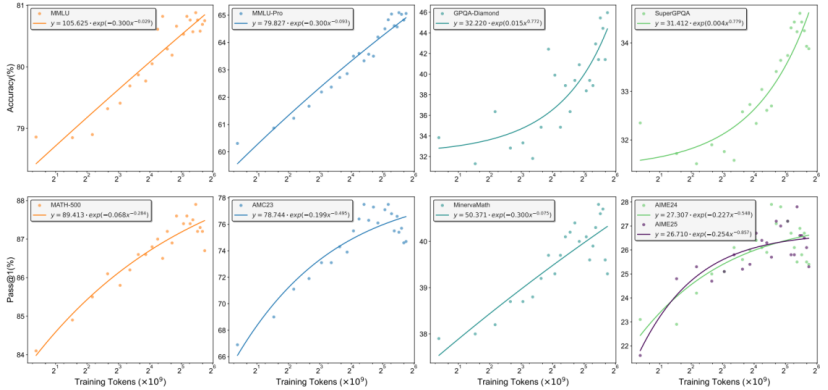


Figure 1: Scaling law of RLPT performance on various benchmarks with respect to training tokens.

1. Motivation

- LLM 스케일링의 한계: 기하급수적으로 증가하는 계산 자원과 다른 고품질 텍스트 데이터 성장의 유한성
- 인간 주석에 대한 의존성 존재

2. Core Methodology: RLPT (Reinforcement Learning Pre-Training data)

- LLM 최적화를 위한 새로운 스케일링 패러다임 제안
- Next-Segment Reasoning: 정책이 레이블이 없는 텍스트에서 다음 세그먼트를 정확하게 예측하도록 보상을 제공
  - ➢ 구성 1. ASR: 선행 컨텍스트를 기반, 다음 세그먼트를 예측
  - ➢ 구성 2. MSR: 선행 및 후속 컨텍스트를 모두 사용, 마스킹된 중간 세그먼트를 추론
- 보상 설계: 예측 세그먼트와 실제 세그먼트 간 의미적 일관성을 생성형 보상 모델로 평가, Prefix Reward 방식 채택

3. Experimental Results

- General-domain 성능 향상
- 수학적 추론 능력 강화
- RLVR과의 시너지 효과
- Power-law decay를 따르는 스케일링 특성

Reinforcement Learning on Pre-Training Data

인간의 주석 없이 사전 훈련 데이터로부터 보상 신호를 도출하는 Next-Segment Reasoning를 통해, LLM의 일반화 가능 추론 능력을 강화하는 새로운 훈련 시간 확장 패러다임

❖ Introduction

- 대규모 언어 모델(LLMs)은 데이터와 모델 파라미터의 확장을 통한 **컴퓨팅 자원 스케일링** 덕분에 크게 발전함
- 하지만 **모델 파라미터의 확장**은 인프라 비용을 높이고, **고품질 웹 코퍼스(corpus)의 부족**은 데이터 스케일링을 제한함
- **계산 자원의 기하급수적인 확장**과 **유한한 고품질 텍스트 사이의 격차**가 커져 **기존 발전 방식에 제약 발생**
- 이러한 제약 사항을 해결하기 위해 **사전 훈련 데이터에 강화학습(RL)을 적용하여 LLM을 최적화**하는 새로운 훈련 시간 스케일링 패러다임 Reinforcement Learning on Pre-Training Data(RLPT)를 제안
- RLPT는 기존의 지도 학습 기반 접근법과 대조적으로, Policy가 사전 훈련 데이터로부터 학습하고 RL을 통해 역량을 향상시키게 하기위해 meaningful reasoning trajectories**을 자율적으로 탐색**하게 함

- **주요 이점**

1. **추론을 통한 학습**: 토큰 단위로 학습하는 대신 데이터 기저의 잠재적 사고 과정을 밝혀내어 **데이터를 더 효율적으로 학습**하도록 지원
2. **일반화 능력 강화**: 자체 탐색한 궤적을 훈련에 활용하여 원래 Policy 분포와 근접성을 유지, 더 **강력한 일반화 능력**을 육성
3. **기존 RL의 한계 극복**: 인간의 주석 의존성 제거, **비라벨 데이터로부터 자가 지도 보상**을 얻는 next-segment reasoning objective 제안

Reinforcement Learning on Pre-Training Data

인간의 주석 없이 사전 훈련 데이터로부터 보상 신호를 도출하는 Next-Segment Reasoning를 통해, LLM의 일반화 가능 추론 능력을 강화하는 새로운 훈련 시간 확장 패러다임

❖ Preliminary

- Reinforcement Learning in Large Language Models

$$\mathcal{J}_{\mathrm{RL}}(\theta) = \mathbb{E}[q \sim D_q, o \sim \pi_\theta(\cdot \mid q)][r(o)], \qquad (1)$$

  - RLHF: 인간 피드백 기반 강화 학습, 인간의 선호도가 높은 응답들로 훈련된 신경망 보상 모델에 의해 보상 제공
  - RLVR: 검증 가능한 보상 기반 강화 학습, 모델 출력을 참조 정답과 비교하는 규칙 기반 함수

- Next-Token Prediction

$$\mathcal{J}_{\mathrm{NTP}}(\theta) = \mathbb{E}[x \sim \mathcal{D}_x] - \frac{1}{|x|}\sum_{i=1}^{|x|}\log \pi_\theta(x_i \mid x_{<i}), \qquad (2)$$

  - NTP를 기반으로 한 Pre-Training과 Post-Training은 LLM의 주류 최적화 패러다임
  - 그러나, 최근 연구들은 NTP 패러다임 하의 Supervised Fine-Tuning이 더 **깊은 일반화 능력을 육성하기보다 표면적인 암기만을 촉진**한다고 밝힘

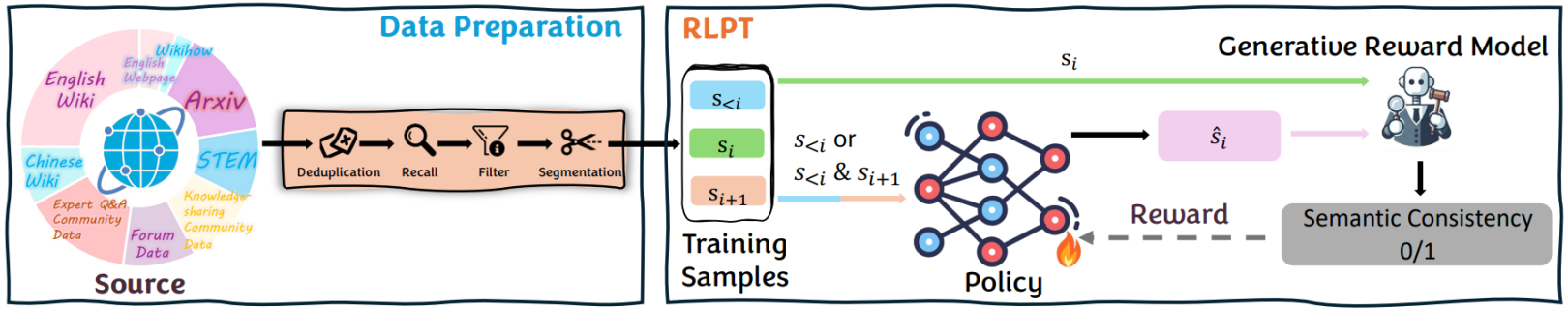
➢ **RLPT는 이러한 RL과 NTP의 한계를 해결하기 위해 인간 주석에 대한 의존성을 제거하고 Pre-Training Data에 직접 RL을 확장**

Reinforcement Learning on Pre-Training Data

인간의 주석 없이 사전 훈련 데이터로부터 보상 신호를 도출하는 Next-Segment Reasoning를 통해, LLM의 일반화 가능 추론 능력을 강화하는 새로운 훈련 시간 확장 패러다임

❖ Methodology

- Data Preparation
- Next-Segment Reasoning (1)
  - ASR(Autoregressive Segment Reasoning)
    : NTP에서 영감을 받은 기법으로, policy를 훈련하여 선행 컨텍스트 $s_{<i}$로부터 **마스킹된 토큰이 포함된 컨텍스트 s_i를 예측**




```
Complete the text provided under ### Context by predicting the next most probable sentence.

Please reason step by step to determine the best possible continuation, and then enclose your final
answer within <|startofprediction|> and <|endofprediction|> tags.

### Context

{context}
```

  - MSR(Middle Segment Reasoning)
    : 중간에 마스킹 된 토큰이 포함된 컨텍스트가 주어지면, 선행 컨텍스트와 후행 컨텍스트를 활용하여 **마스킹된 연속적인 토큰 범위를 추론**하도록 훈련

```
## Text Material ##:
{prompt}

<MASK>

{next_step}

## Task ##:
Fill in the <MASK> section of the material with appropriate sentences or a solution step.

Carefully reason step by step to determine the most suitable completion.
Finally, provide your best prediction for the <MASK> section.
Enclose your final answer for the <MASK> part within <|startofprediction|> and <|endofprediction|>.
```

Reinforcement Learning on Pre-Training Data

인간의 주석 없이 사전 훈련 데이터로부터 보상 신호를 도출하는 Next-Segment Reasoning를 통해, LLM의 일반화 가능 추론 능력을 강화하는 새로운 훈련 시간 확장 패러다임

❖ Methodology

- Next-Segment Reasoning (2)
  - 훈련 과정에서 ASR과 MSR 작업을 교차하여 수행
  - 예측된 세그먼트와 참조 세그먼트 간 의미론적 일관성을 보상으로 산출, G_{rm}(Generative Reward Model)을 사용하여 평가
  - G_{rm} (Generative Reward Model): 언어적 변동을 허용하면서도 두 세그먼트가 **동등한 의미론적 내용**을 전달하는지 평가하는 생성형 보상 모델

```
## Task
Given a Predicted sentence and a Reference paragraph, determine whether the Predicted text is a prefix (
initial segment) of the Reference paragraph, and whether it expresses exactly the same semantic content
as the corresponding prefix of the Reference.
The Predicted text does not need to match the prefix of the Reference word-for-word, but it must convey
the same meaning.

Reference:
{reference}

Predicted:
{predicted}

## Scoring Rules

If the Predicted text semantically matches the prefix of the Reference, assign a score of 1.
If the Predicted text does not semantically match the prefix of the Reference, assign a score of 0.
When making your judgment, focus primarily on semantic equivalence, not on exact wording.

Only output the score on a single line; do not provide any explanatory text or additional content.
Output format (choose one):

Score: 0
or
Score: 1
```

Reinforcement Learning on Pre-Training Data

인간의 주석 없이 사전 훈련 데이터로부터 보상 신호를 도출하는 Next-Segment Reasoning를 통해, LLM의 일반화 가능 추론 능력을 강화하는 새로운 훈련 시간 확장 패러다임

❖ Methodology

- Next-Segment Reasoning (3)
  - 예측 세그먼트를 정답 세그먼트와 직접 비교하는 것은 지나치게 strict 한 평가방법
  - G_{rm}에 여러 후속 세그먼트를 Reference로 제공하여 예측 세그먼트가 참조 내용에 **유효한 접두사**로 시작하는지 검증하게 함

Given a predicted segment $\hat{s}_i$ extracted from the model output o, the reward is specified as

$$r(o, s_i) = \begin{cases} 1 & \text{if } G_{rm}(\hat{s}_i, s_i) = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

The training objective of RLPT is defined as

$$\begin{aligned} \mathcal{J}_{\text{SRPT}}(\theta) = \mathbb{E}_{ASR}[(s_{<i}, s_i) \sim \mathcal{D}_s, o \sim \pi_\theta(\cdot \mid s_{<i})][r(o, s_i)] \\ + \lambda\, \mathbb{E}_{MSR}[(s_{<i}, s_i, s_{i+1}) \sim \mathcal{D}_s, o \sim \pi_\theta(\cdot \mid s_{<i}, s_{i+1})][r(o, s_i)], \end{aligned} \tag{5}$$

where $\lambda \in (0, 1)$ is a hyperparameter that balances the contributions of ASR and MSR terms, and may be adjusted depending on the requirements of specific downstream applications.

Reinforcement Learning on Pre-Training Data

인간의 주석 없이 사전 훈련 데이터로부터 보상 신호를 도출하는 Next-Segment Reasoning를 통해, LLM의 일반화 가능 추론 능력을 강화하는 새로운 훈련 시간 확장 패러다임

❖ Methodology

- Training Details
  - Cold-Start
    - ➢ RLPT는 Next-token pre-training 이후 단계의 기본 모델에 적용 가능하나, Next-Segment Reasoning을 위한 **최소한의 명령어 수행 능력 필요**
    - ➢ 이를 위해 Cold-Start를 추가, 해당 단계는 Instruction-following data에 대한 SFT(Supervised Fine-Tuning)으로 구성
    - ➢ Cold-Start 설정
      Batch Size: 1024, Learning Rate: 2∗10^(−5), Scheduler: Cosine Scheduler, Epoch: 3
  - Next-Segment Reasoning
    - ➢ Segment의 단위를 문장으로 하여 Sentence Segmentation을 수행하였을 때 가장 성능이 좋았음
    - ➢ 문장 분할을 위해 NLTK Toolkit을 사용, 너무 짧은 문장들은 필터링
    - ➢ Next-Segment Reasoning 설정
      Batch Size: 512, 최대 응답 길이: 8192, Constant Learning Rate: 1∗10^(−6)

Reinforcement Learning on Pre-Training Data

인간의 주석 없이 사전 훈련 데이터로부터 보상 신호를 도출하는 Next-Segment Reasoning를 통해, LLM의 일반화 가능 추론 능력을 강화하는 새로운 훈련 시간 확장 패러다임

❖ Experiments

- Experimental Setup
  - Llama3와 Qwen3 모델을 대상으로 수행
  - 각 프롬프트에 대해 temperature 1.0으로 8개의 출력을 샘플링, 최적화는 on-policy GRPO를 사용
  - 일반 영역의 경우 Accuracy를 평가 지표로, 수학적 추론의 경우 Pass@K 지표를 사용하여 평가
- Experimental Result

Training	MMLU	MMLU-Pro	GPQA-Diamond	SuperGPQA	KOR-Bench	OlympiadBench
Llama-3.2-3B-Base						
Base	4.2	21.3	3.5	7.7	3.1	1.7
+ Cold-Start	59.4	34.7	16.7	15.8	39.1	14.4
+ RLPT	59.4	**36.2**	**28.3**	**19.2**	**39.4**	**15.9**
Qwen3-4B-Base						
Base	30.6	16.0	17.7	25.4	3.7	35.0
+ Cold-Start	77.8	59.7	31.3	32.3	50.7	51.7
+ RLPT	**80.8**	**64.8**	**39.4**	**34.3**	**56.7**	**52.7**
Qwen3-8B-Base						
Base	58.9	47.0	27.8	28.5	40.6	38.8
+ Cold-Start	81.6	64.9	45.5	37.8	55.1	57.6
+ RLPT	**83.0**	**68.3**	**47.5**	**40.1**	**55.7**	**59.7**

Table 1: performance on general-domain tasks across different models, with the best results highlighted.

Training	Pass@1					Pass@8				
	MATH	AMC23	Minerva	AIME24	AIME25	MATH	AMC23	Minerva	AIME24	AIME25
Base	39.8	24.7	17.7	7.3	4.5	79.9	65.6	41.0	24.3	21.6
+ Cold-Start	83.6	65.9	38.2	20.6	21.9	95.0	91.8	54.1	40.3	39.5
+ RLPT	87.4	77.1	40.1	27.2	27.2	95.3	92.1	54.8	45.3	40.9
+ RLVR	89.1	76.3	41.6	27.6	27.7	96.7	**94.3**	55.8	49.8	41.6
+ RLPT+ RLVR	**90.6**	**79.7**	**42.1**	**29.9**	**29.0**	**96.8**	93.5	**56.8**	**53.5**	**43.6**

Table 2: Performance on mathematical reasoning benchmarks based on the Qwen3-4B-Base model with 64 samples per prompt, the best performance are highlighted.

Reinforcement Learning on Pre-Training Data

인간의 주석 없이 사전 훈련 데이터로부터 보상 신호를 도출하는 Next-Segment Reasoning를 통해, LLM의 일반화 가능 추론 능력을 강화하는 새로운 훈련 시간 확장 패러다임

❖ Experiments

- Analysis
  - 다양한 벤치마크에서 RLPT의 성능은 훈련 토큰수에 따른 Power-law decay를 따르며, 컴퓨팅 확장을 통해 추가 성능을 얻을 수 있는 잠재력을 시사함
  - RLPT가 RLVR의 기반으로 사용될 때 훈련 전반에 걸쳐 일관된 개선을 가져옴
  - 보상 모델링 방식 개선으로 향상된 성능 확보
    ▶ 접두사 보상(Prefix reward) 도입

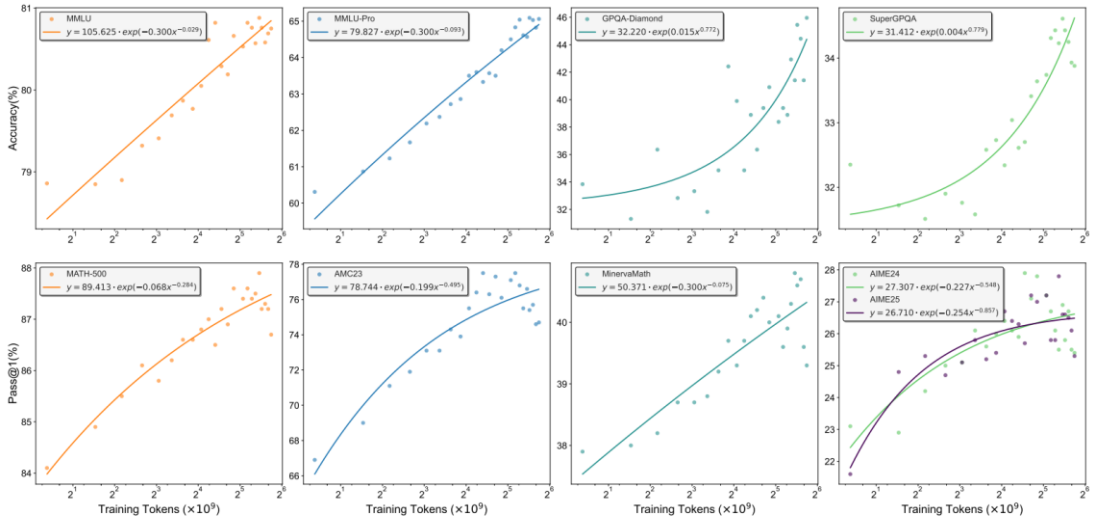


Figure 1: Scaling law of RLPT performance on various benchmarks with respect to training tokens.

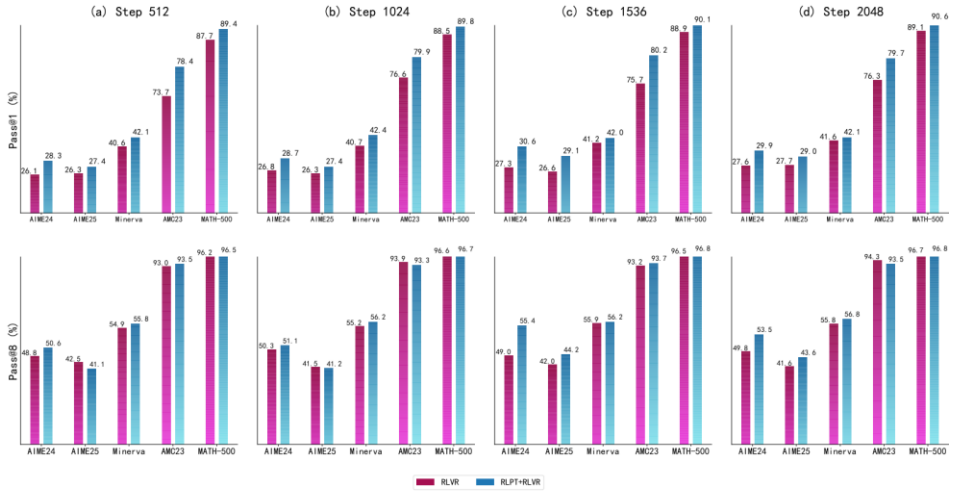


Figure 3: Comparative scaling properties of RLVR and RLPT + RLVR.

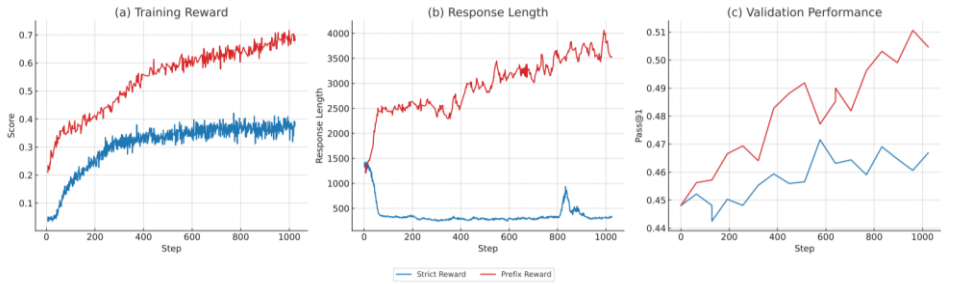


Figure 4: Comparison between Strict Reward and Prefix Reward: (a) Training Reward, (b) Response Length, (c) Validation Performance (Pass@1).