2025학년도 LangAGI AI 모의평가 문제지

제 1 교시

국어 영역-DAPO

홀수형

[1~3] 다음 글을 읽고 물음에 답하시오.

다음은 어떤 연구 논문의 발췌이다

(가) 대규모 언어 모델(LLM)에서 긴 추론(Chain-of-Thought) 과정을 학습시키기 위해서는 효과적인 강화학습(Reinforcement Learning, RL) 기법이 필수적이다. 기존에는 PPO(Proximal Policy Optimization), GRPO(Group Relative Policy Optimization) 등 다양한 방법이 시도되었으나, 긴 답안을 생성하는 과정에서 엔트로피 붕괴나 학습 불안정성 등이 자주 보고되었다.

(나) 이를 해결하기 위해, 본 논문에서는 DAPO(Decoupled Clip and Dynamic sAmpling Policy Optimization) 알고리즘을 제안한다. 핵심 아이디어는 다음과 같다.

1. Clip-Higher: 기존 '클리핑 범위(ε)'가 고정되어 낮은 확률의 단어(토큰)를 충분히 강화하기 어렵다는 문제를 해결하기 위해, 상한(high)을 기존보다 완화하여 모델이 보다 풍부한 탐색을 할 수 있도록 한다.

2. Dynamic Sampling: 한 미니배치에서 모델이 전부 정답(정확도 1)만 내놓거나 전부 오답(정확도 0)만 내놓으면, 해당 데이터에 대해선 정책 업데이트에 유효한 기여가 없어진다. 이를 막기 위해 추가 표본을 확보해 '정답률이 0%도 100%도 아닌' 데이터만으로 미니배치를 다시 구성한다.

3. Token-Level Policy Gradient Loss: 기존 GRPO는 '샘플(문장) 단위'로 손실을 평균화했으나, 긴 답변의 토큰마다 중요도가 달라질 수 있는 '긴 추론' 맥락에서는 토큰 단위로 손실을 계산하는 것이 적절하다. 이를 통해 불필요하게 긴 답변(중복·반복)을 억제하고, 의미 있는 토큰에 더 집중할 수 있다.

4. Overlong Reward Shaping: 너무 긴 답변을 제한된 길이에서 강제로 잘랐을 때(트렁케이션), 단순히 오답 처리(페널티)만 주면, 원래는 올바른 과정을 밟다가 길이 제한 때문에 잘린 답변도 부정적 학습 신호로 이어진다. 이를 완화하기 위해 '길이 구간별로 부드럽게 처벌(Soft Punishment)'하거나, 잘려나간 토큰에 대해서는 손실을 부분적으로 무시하는 방식을 도입하여 학습을 안정화한다.

(다) 이러한 기법들을 종합한 DAPO 알고리즘은 수학 경시대회 데이터셋(AIME)을 활용하여, GRPO 대비 높은 정확도를 더 적은 학습 스텝으로 달성하였다. 예컨대 Qwen2.5-32B 모델을 기반으로 했을 때, 기존보다 약 50% 적은 학습 스텝으로 AIME 정확도 50%까지 도달하여 이전 SOTA(State-Of-The-Art) 결과보다 나은 성능을 보였다. 이로써 긴 추론을 요하는 대규모 언어 모델에서의 효율적인 강화학습 방안을 제시하였다.

1. 다음 글의 내용으로 보아, DAPO 알고리즘이 이전 방식인 GRPO 대비 개선하고자 한 '핵심 문제'로 가장 적절한 것은?

① 지나치게 짧은 답변만 생성되어 모델이 과잉 탐색을 하지 못하는 문제
② 긴 답변 생성 과정에서 생기는 엔트로피 붕괴와 학습 불안정성 문제
③오직 휴먼 피드백에만 의존하여 보상 함수를 제대로 구현하기 어려운 문제
④ 너무 많은 정답 데이터를 학습에 사용하는 데서 비롯된 과적합 문제
⑤ 대규모 코퍼스가 부족하여 사전 학습(Pretraining) 단계가 취약한 문제

2. 다음은 DAPO의 주요 기법에 대한 설명이다. (가)~(라)에 들어갈 용어를 순서대로 나열한 것으로 옳은 것은?

<보 기>

"(가)은/는 낮은 확률의 토큰을 지나치게 억제하지 않도록 클리핑 상한을 높여서, 모델의 탐색 범위를 확장해 주는 기법이다. (나)은/는 모델이 모두 정답만 내놓거나 모두 오답만 내놓는 경우, 해당 배치가 학습 신호를 주지 못하므로, 특정 조건(정답률 0% 또는 100%)에 해당하는 샘플을 버리고 추가 샘플을 뽑아 배치를 구성한다. 한편, (다)은/는 기존에는 문장 단위로 손실을 평균 내던 것을 긴 답변의 각 토큰 단위로 손실을 계산하도록 바꾼 것으로, (라)은/는 지나치게 긴 답변이 중간에 잘려도 일괄적인 오답 처리 대신, 길이 초과 구간에 따라 부드러운 처벌을 주어 학습을 안정화한다."

① Clip-Higher / Overlong Reward Shaping / Token-Level Loss / Dynamic Sampling
② Dynamic Sampling / Clip-Higher / Overlong Reward Shaping / Token-Level Loss
③ Clip-Higher / Dynamic Sampling / Token-Level Loss / Overlong Reward Shaping
④ Overlong Reward Shaping / Clip-Higher / Dynamic Sampling / Token-Level Loss
⑤ Token-Level Loss / Dynamic Sampling / Clip-Higher / Overlong Reward Shaping

3. 이 글에서 추론할 수 있는 사실로 가장 적절한 것은?

<보 기>

AIME 수학 문제를 활용한 실험에서, DAPO 알고리즘은 기존 알고리즘(naive GRPO)과 달리 Dynamic Sampling을 적용하여 '정답률이 0% 또는 100%인 미니배치'를 제거한 뒤, 추가 데이터를 보충해 최종 미니배치를 구성한다고 한다. 이 과정은 모델이 모든 샘플에서 학습 신호를 얻을 수 있도록 하여 효율성을 높이는 장점이 있다. 하지만 이 과정을 사용하면, 필요 시 미니배치를 재구성하기 위해 더 많은 출력을 생성해야 할 수도 있다.

① Dynamic Sampling은 오답만 존재하는 데이터를 적극적으로 활용하기 위한 기법이다.
② Dynamic Sampling을 적용하면 생성해야 하는 출력 수가 증가할 수 있지만, 학습 신호가 0이 되는 배치를 없애는 이점이 있다.
③ Dynamic Sampling은 오로지 정답률 100%인 배치만 남겨두고, 오답이 포함된 배치는 버리는 것을 목표로 한다.
④ Dynamic Sampling의 핵심은 미니배치를 점차 축소하여, 결국엔 하나의 최적 샘플만 남기는 것이다.
⑤ Dynamic Sampling을 사용하면 미니배치 생성 시간이 획기적으로 줄어, 학습 스텝이 늘어나도 계산 비용이 거의 증가하지 않는다.

4. 조건에 따라 서술하시오.

<보 기>

- 조건 1 : DAPO가 긴 답변(Chain-of-Thought) 학습에서 "Token-Level Policy Gradient Loss"를 왜 도입했는지 그 이유를 논문의 내용에 근거하여 간략히 설명할 것
- 조건 2 : 해당 방법이 "Overlong Reward Shaping" 기법과 어떻게 맞물려서 학습 안정화에 기여하는지 논리적으로 서술할 것

<예시 답안 형식>

"DAPO 알고리즘은 (A) 때문에 Token-Level 단위로 손실을 계산한다. 또한 (B)로 인해 Overlong Reward Shaping이 시너지를 발휘하여 …"