**--enable\_hybrid\_engine이 안되는 에러**

issue에서 찾아보니 deepspeed가 hybrid engine을 업데이트 했지만 deepspeedchat이 아직 이 업데이트를 반영하지 않아서 에러가 발생.

Deepspeed 버전을 0.9.5에서 0.9.0으로 변경하여 해결

**옵션에 따른 VRAM 변화량**

--enable\_hybrid\_engine을 작동시키니 VRAM 소모량이 10GB로 증가함

--only\_optimize\_lora 대신에 Gradinet\_checkpointing을 사용하니 7GB 사용

--enable\_hybrid\_engine을 끄고, Gradinet\_checkpoing을 사용하니 속도는 좀 느려졌지만 VRAM 6.4GB 사용

-- enable\_ema()를 추가하면 어떻게 되는지 알아보기

**--enable\_ema()**

Actor 모델 가중치의 지수 이동 평균을 수집하기 위해서 다른 모델을 추가한다.

**KL divergence가 너무 작아서 모델 학습이 이상하게 된 case**

<https://github.com/microsoft/DeepSpeedExamples/issues/586>

**Policy Gradient**

보상의 기대값을 최대화하는 방향으로 가중치를 조정하는 강화학습 알고리즘

Actor Loss:

주어진 학습 prompt에 대해서 정책을 학습한다.

언어 모델이 어떤 답변을 생성할지에 대한 확률적인 분포를 학습

주어진 prompt에서 생성된 응답의 log probability와 이전에 생성된 응답의 log probability의 확률 차이를 계산하고 policy gradient알고리즘을 통해 이 차이를 계산하여 언어 모델을 업데이트한다. 업데이트를 할 때, critic 모델에서 계산 값을 사용한 advantage (현재 응답의 상대적인 품질을 나타내는 값)

**Actor 모델**

**Critic 모델**

Critic 모델은 현재 상태의 가치를 평가한다. 즉, 시퀀스의 각 토큰 위치에서 현재 상태가 얼마나 좋은지를 예측. Reward 모델은 보상을 최대화 하지만, critic 모델은 보상을 예측하는데 사용된다.

**Reward 모델**

Actor 모델이 만든 시퀀스에 대한 보상 점수를 평가.

**Advantage:**

텍스트, 스크린샷, 폰트, 라인이(가) 표시된 사진

자동 생성된 설명

rewards[:, t]: 시간 t에서의 보상값(reward) (즉시 보상)

gamma: 할인 인자(discount factor), 0 <= gamma <= 1

nextvalues: 시간 t+1에서의 상태값(state value) (다음 상태의 기대 리턴 값)

values[:, t]: 시간 t에서의 상태값(state value) (현재 상태의 기대 리턴 값)

lambda: GAE(Genaralized Advantage Estimation)에서 사용되는 Generalized Advantage Estimation 계수, 0 <= lambda <= 1



values[:, start:]: PPO 에이전트가 현재 시점(start) 이후의 기대 리턴 값을 나타냅니다. 즉, 미래의 리턴값을 계산합니다.