**최종 모델 학습을 위한 데이터셋**

최종 모델:

Ko-Alpaca (Stanford Alpaca 데이터셋, 네이버 지식IN 베스트 질문)

KULLM (GPT4ALL, Databricks-Dolly, Vicuna)

우리는 최종 학습 instruction으로 Best Open Assistant dataset을 사용할 예정

텍스트, 전자제품, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

**Reward 모델을 학습에 사용할 foundation models**

현재 사용이 가능한 가장 강력한 한국어 모델들로 사용할 예정

OpenAI ChatGPT

Google Bard

Ko-Alpaca

**Reward 모델의 학습방법 및 과정 이해하기**

데이터셋: 하나의 prompt에 해당하는 여러 모델의 답변을 사람이 직접 순위를 매겨서 제작함 (Comparison dataset)

이 데이터셋에서 RM을 학습시켜 Labeler가 선호하는 모델 출력을 예측한다.

RM을 PPO 알고리즘의 reward function으로 사용하여 모델을 학습시키고 reward를 최대화한다.

기존 RLHF는 광범위한 Task에서 동작이 크게 개선되지만, 안전성과 신뢰성은 여전히 부족하다 따라서 우리는 안전성을 개선한다.

텍스트, 영수증, 폰트, 화이트이(가) 표시된 사진

자동 생성된 설명

Step1) SFT 데이터셋(prompt-output pair dataset) 구축

Step2) SFT 데이터셋에서 여러 모델들의 답변과 ranking이 추가된 RM 데이터셋 생성

Step3) SFT데이터셋에서 prompt만 가져와서 구축

SFT 모델을 선정하고 instruction tuning 진행

SFT 모델들의 unembedding layer를 제거하고 output의 ranking을 label로 하여 cross-entropy loss로 학습을 진행. Validation set의 RM점수에 따라서 최종 SFT 모델을 선정

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

Label 가이드 라인

유용하고, 정직하며, 무해한 특징을 가지도록 평가한다.

유용 > 정직 > 무해