모델선택 및 데이터셋 조사

1. CanAICode Leaderboard

* 사람이 만든 인터뷰 질문들이 주어지고 AI가 답을 한다.
* Junior-dev: 여러 언어로 된 테스트 케이스들의 모음집이다.
* Humaneval: OpenAI에서 만든 파이썬 언어로 된 164개의 테스트케이스 모음집이다.
  + OpenAI의 Codex 모델을 평가한 테스트케이스가 Humaneval이다.

1. LLaMa2가 Code generation에서 성능이 안 좋은지

codeCherryPop도 CanAICode의 평가 점수를 받기 위해 준비중이다. CanAICode 깃헙 레포에 보니 평가에 필요한 프롬프트가 올라와 있다.

점수를 확인해본 뒤에 다른 모델들과 비교해 보고 LLaMa2로 갈지 결정 하는 게 좋겠다.

다만 평가 점수의 순위가 낮더라도 7B 모델에 12만개의 데이터로 finetuning된 모델이라는 점을 고려해서 비슷한 크기의 모델들과 비교해야 한다. 13B로 사이즈를 키우고 데이터셋을 늘리면 성능이 더 좋아질 수 있다.

* Backbone 모델: llama2 7B HF
* 사용한 데이터셋: sahil2801/code\_instructions\_120k
* 프롬프트 및 미세튜닝 스타일: Alpaca Style instruction tuning
  + llama2 style[[1]](https://github.com/the-crypt-keeper/can-ai-code/blob/main/prompts/llama2-chat.txt) - [INST]<<SYS>>
  + Alpaca style[[2]](https://github.com/the-crypt-keeper/can-ai-code/blob/main/prompts/Alpaca.txt)[[3]](https://github.com/the-crypt-keeper/can-ai-code/blob/main/prompts/Alpaca-Input.txt) - ### Instruction: ### Input: ### Response:
* 훈련 방식: QLoRa

자세한 훈련 방식을 참고하려 했으나 해당 유저가 올린 노트북 링크가 열리지 않아서 유튜브 댓글로 요청했다.

나온 지 얼마 되지 않아서 관련 정보가 적다.

1. Java instruct instruction-tuning 데이터 셋이 있는지