



LLM_(4)

말 잘 듣는 모델 만들기

말 잘 듣는 모델 만들기

- GPT-3 를 Chat GPT 로 만들기
 - GPT3
 - 학습 방법 : 다음 나올 단어를 예측하는 방식
 - 장점 및 기능 : 사용자의 말에 이어질 법한 텍스트를 잘 생성
 - 한계 : 요청을 파악하고 대답을 해주는 개념이 아니다



말 잘 듣는 모델 만들기

- GPT-3 를 Chat GPT 로 만들기
 - GPT3
 - 학습 방법 : 다음 나올 단어를 예측하는 방식
 - 장점 및 기능 : 사용자의 말에 이어질 법한 텍스트를 잘 생성
 - 한계 : 요청을 파악하고 대답을 해주는 개념이 아니다
- Chat GPT로 변화시킬 수 있는 방법
 - 1단계 지시 데이터 셋
 - 질문(요청), 답변 형식의 데이터 셋을 생성
 - GPT-3가 위 데이터 셋을 기반으로, 사용자 요청에 대해 응답할 수 있도록 학습
 - 2단계
 - 사용자가 더 선호하는 답변을 생성할 수 있도록 추가 학습
 - 방법1. 강화학습 사용 O (예 : RLHF, PPO)
 - 방법2. 강화학습 사용 X (예 : 기각 샘플링, DPO)

말 잘 듣는 모델 만들기

- GPT-3 를 Chat GPT 로 만들기
 - GPT3
 - 학습 방법 : 다음 나올 단어를 예측하는 방식
 - 장점 및 기능 : 사용자의 말에 이어질 법한 텍스트를 잘 생성
 - 한계 : 요청을 파악하고 대답을 해주는 개념이 아니다
- Chat GPT로 변화시킬 수 있는 방법
 - 1단계
 - 질문(요청), 답변 형식의 데이터 셋을 생성
 - GPT-3가 위 데이터 셋을 기반으로, 사용자 요청에 대해 응답할 수 있도록 학습
 - 2단계
 - 사용자가 더 선호하는 답변을 생성할 수 있도록 추가 학습
 - 방법1. 강화학습 사용 O (예 : RLHF, PPO)
 - 방법2. 강화학습 사용 X (예 : 기각 샘플링, DPO)

말 잘 듣는 모델 만들기

- Contents

- ❖ 사전 학습과 지도 미세 조정
- ❖ 선호도 반영하는 강화학습
- ❖ 선호도 반영하는 비강화학습



- 들어가기 전
 - 목적 : LLM이 사용자의 요청에 응답할 수 있도록 학습하는 방법을 알아보기
 - 예시 : 코딩 테스트 서비스를 이용하여 사용자가 어떤 문제를 시작해야 하고 코딩 테스트를 통과할 수 있을지



사전 학습과 지도 미세 조정

- 사전학습

- 코딩 개념 익히기 :

- 파이썬이라는 프로그래밍 언어 공부
기본적인 문법, 정의, 개념, 사용 방법 등 공부
자료구조, 알고리즘 책 공부

- LLM 사전 학습 :

- 인터넷에 있는 다양한 텍스트 데이터로 사전 학

Recall) LLM : 딥러닝 기반의 언어 모델

다음 단어를 예측하는 언어 모델링을 통해 텍스트를
이해하는 방법을 학습한다.



사전 학습과 지도 미세 조정

- 사전학습

- LLM 사전 학습 :

- 인터넷에 있는 다양한 텍스트 데이터로 사전 학습

- Llama-2 모델 : 10TB 텍스트 데이터
(코드, 블로그, 기사, 광고 ...)

- 다음 단어를 예측하는 방법으로 학습
 - LLM이 특정한 형태로 응답하거나 요청에 따라 응답하기 어려움
 - 단지, 언어에 대한 전체적인 이해도와 예측 성능 증가

Recall) 사전학습을 통해
GPT-3 가 다음 올 단어를 잘 예측하는 상황

사전 학습과 지도 미세 조정

- 지도 미세 조정
 - 코딩 개념 익히기 -> 코딩 테스트 연습 문제 풀이 :
간단한 예제를 충분히 풀어보기
자주 나오는 문제와 정답을 보고 풀이 과정 익히기
주어진 문제에 맞게 코드를 작성하는 법 연습
 - LLM 사전 학습 -> 지도 미세 조정 :
사용자의 요청을 적절히 해석하기
응답의 형태를 적절히 작성하기
요청과 응답이 잘 연결되도록 학습하기

지도 : Supervised, 학습 데이터에 정답이 포함

정렬 : Alignment, 지도 미세 조정을 통해 LLM이 사용자의 요청에 맞춰
응답하도록 학습



사전 학습과 지도 미세 조정

- 지도 미세 조정

- LLM 사전 학습 -> 지도 미세 조정 :
 사용자의 요청을 적절히 해석하기
 응답의 형태를 적절히 작성하기
 요청과 응답이 잘 연결되도록 학습하기

- 지도 미세 조정에 사용하는 데이터셋
 지시 데이터셋 : 사용자의 지시에 맞춰 응답한 데이터셋

지식 데이터셋 << 사전 학습데이터셋
(형식의 다양성)
(양질의 데이터)

Recall) 딥러닝 모델은 학습 데이터에 있는 행동을 배우기 때문에,
학습 데이터가 적다면 그 행동은 잘 배우지 못한다.



사전 학습과 지도 미세 조정

- 지도 미세 조정 문제 개선안

한계점 : 지시 데이터 셋 << 사전 학습데이터 셋
(형식의 다양성)
(양질의 데이터)

개선안 : 사용자의 요구사항과 이에 대한 응답을 구조화한 데이터 구축

2022년 Open AI의 Chat GPT개발 당시,
위 한계점 극복한 방법 :

Recall) 딥러닝 모델은 학습 데이터에 있는 행동을 배우기 때문에,
학습 데이터가 적다면 그 행동은 잘 배우지 못한다.

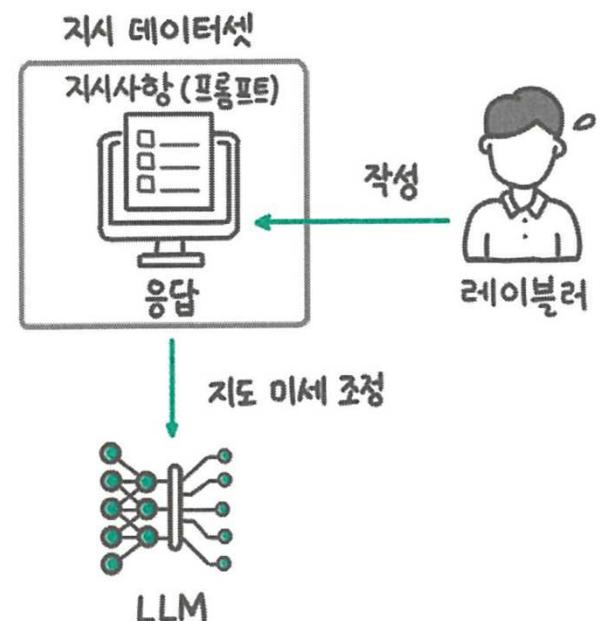
사전 학습과 지도 미세 조정

- 지도 미세 조정 문제 개선안

한계점 : 지시 데이터셋 << 사전 학습데이터셋
(형식의 다양성)
(양질의 데이터)

개선안 : 사용자의 요구사항과 이에 대한 응답을 구조화한 데이터 구축

2022년 Open AI의 Chat GPT개발 당시,
위 한계점 극복한 방법 : 레이블러 고용



사전 학습과 지도 미세 조정

- 지도 미세 조정_지시 데이터셋

지시 데이터셋 형태

- Instruct (지시사항) : 사용자의 요구사항을 표현한 문장
- Input (입력) : 답변을 하는 데 필요한 데이터
- Output (출력) : 지시사항과 입력을 바탕으로 한 정답 응답
- text : 지시사항, 입력, 출력을 정해진 포맷으로 하나로 묶은 데이터

▼ 예제 4.1 알파카 데이터 형태(출처: <https://huggingface.co/datasets/tatsu-lab/alpaca?row=5>)

```
{  
    "instruction": "Create a classification task by clustering the given list of items.",  
    "input": "Apples, oranges, bananas, strawberries, pineapples",  
    "output": "Class 1: Apples, Oranges\nClass 2: Bananas, Strawberries\nClass 3:  
        \"text\": \"Below is an instruction that describes a task, paired with an input  
        that provides further context. Write a response that appropriately completes the  
        request.\n\n### Instruction:\nCreate a classification task by clustering the given  
list of items.\n\n### Input:\nApples, oranges, bananas, strawberries, pineapples\n\n### Response:\nClass 1: Apples, Oranges\nClass 2: Bananas, Strawberries\nClass 3:  
Pineapples\",  
}
```

사전 학습과 지도 미세 조정

- 지도 미세 조정_지시 데이터셋

지시 데이터셋 형태

- Instruct (지시사항) : 사용자의 요구사항을 표현한 문장
- Input (입력) : 답변을 하는 데 필요한 데이터
- Output (출력) : 지시사항과 입력을 바탕으로 한 정답 응답
- text : 지시사항, 입력, 출력을 정해진 포맷으로 하나로 묶은 데이터

▼ 예제 4.2 지시 데이터셋의 형식을 갖추기 위한 알파카 템플릿

```
f"""
```

Below is an instruction that describes a task, paired with an input that provides further context.

Write a response that appropriately completes the request.\n\n

```
### Instruction:\n{instruction}\n\n
```

```
### Input:\n{input}\n\n
```

```
### Response:\n{output}
```

```
"""
```

사전 학습과 지도 미세 조정

- 지도 미세 조정_지시 데이터셋

지시 데이터셋을 LLM은 어떻게 학습할까?

- 사전학습과 동일
- 다음 단어 예측하는 인과적 언어 모델링 사용

지도 미세 조정 :

- LLM이 학습하는 방식은 동일
- 학습하는 데이터셋에 차이점 존재

지시 데이터셋 구성 차이

→ LLM 성능 결정



사전 학습과 지도 미세 조정

- 좋은 지시 데이터셋 (By Meta, MS. 2023)

지시 데이터셋의 양 :

Less Is More for Alignment, 2023, Meta 에서

파라미터가 650억개인 LLaMa모델을 정렬하는데 선별한 지시 데이터셋

→ 1000개

→ LIMA 모델 제시

성능 : 52,000개의 지시 데이터셋으로 학습한 알파카 보다 양질의 답변
Bard, GPT-4와 비교 시, 40~50% 답변은 LIMA 우세 혹은 비슷

사전 학습과 지도 미세 조정

- 좋은 지시 데이터셋 (By Meta, MS. 2023)

지시 데이터셋의 지시사항 형태 :

Less Is More for Alignment, 2023, Meta 에서

LLaMa모델을 구성하는데 사용 시도한 데이터셋 2가지

1. Wikihow
2. Stack Exchange

→ 두 가지 모두 사용자가 질문하고 다른 사용자가 답변한 형태

Wikihow : 답변 품질 우수 / 질문 형식이 모두 How to로 동일

Stack Exchange : 답변 품질 낮음 / 질문 형식이 다양

→ Stack Exchange 데이터 중 답변의 품질이 좋은 것만 선별

사전 학습과 지도 미세 조정

- 좋은 지시 데이터셋 (By Meta, MS. 2023)
 1. 데이터셋 양은 1000개 정도도 충분
 2. 질문의 형식이 다양하며, 답변의 품질이 높은 데이터 선별하여 사용
- 피상적인 정렬 가설(superficial alignment hypothesis)을 주장
- 피상적인 정렬 가설 :
- 모델의 지식이나 능력은 사전 학습 단계에서 대부분 학습 정렬 데이터에서는 답변의 형식이나 내용의 나열 정도만 추가로 학습하기 때문에 적은 정렬 데이터로도 사용자가 원하는 답변을 생성할 수 있다.

사전 학습과 지도 미세 조정

- 좋은 지시 데이터셋
 1. 데이터셋 양은 1000개 정도도 충분
 2. 질문의 형식이 다양하며, 답변의 품질이 높은 데이터 선별하여 사용

(옵션) 고품질의 데이터 생성하여 추가 학습 가능

EX) phi

스택+ : 공개된 코드 데이터셋 Stack 그대로 사용

코드 텍스트북 : 교육적 가치가 높은 데이터를 선별한 코드 텍스트 북을 학습데이터로 사용

코드 예제 데이터셋 : GPT-3.5로 생성한 고품질의 예제 데이터셋으로 추가학습

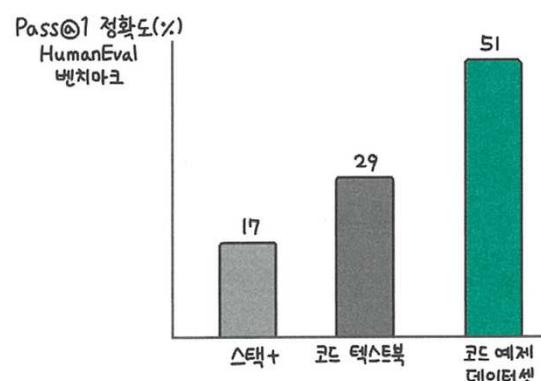


그림 4.7 학습 데이터에 따른 파이 모델의 성능 비교(출처: <https://arxiv.org/abs/2306.11644>)

선호도 반영하는 강화 학습

- 선호 데이터셋

- 선호도 기준을 새우고, 해당 기준을 평가하는 모델 만들기

핵심 : 모델을 학습시킬 학습 데이터를 구축

Ex) 코드 가독성 평가하는 모델

- 가독성 비교 데이터셋 구축

- 두 코드 A, B 비교하여 더 가독성 높은 코드 선택

- 선택한 코드를 선호 데이터, 선택하지 않은 코드를 비선호 데이터

Chat GPT)

- 사전 학습

- 지도 미세 조정, 지시사항 입력

- 여러 답변 생성

- 생성된 답변을 레이블러가 선호도 판단 -> 선호 데이터셋 구축

- 선호 데이터셋을 사용해 리워드 모델이 점수를 부여하는 모델 제작

선호도 반영하는 강화 학습

- 선호 데이터셋

- 선호도 기준을 새우고, 해당 기준을 평가하는 모델 만들기

Chat GPT)

- 사전 학습

- 지도 미세 조정, 지시사항 입력

- 여러 답변 생성

- 생성된 답변을 레이블러가 선호도 판단 -> 선호 데이터셋 구축

- 선호 데이터셋을 사용해 리워드 모델이 점수를 부여하는 모델 제작

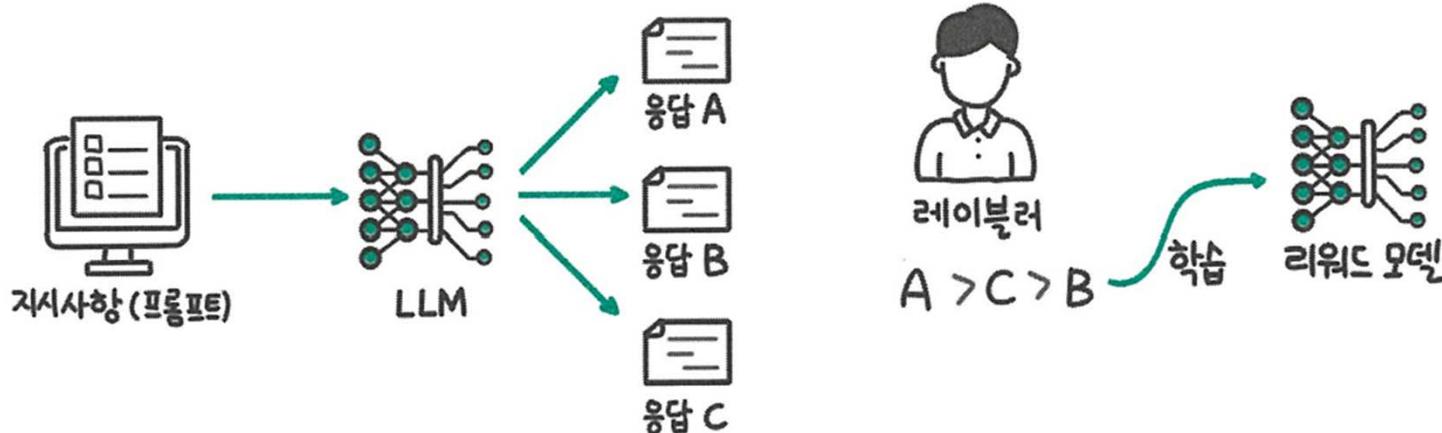


그림 4.10 챗GPT를 학습하는 과정에서의 선호 데이터셋 구축과 리워드 모델 학습

선호도 반영하는 강화 학습

- RLHF

- RLHF (Reinforcement Learning from Human Feedback)

Training language models to follow instructions with human feedback, 2022
, OpenAI에서 강화 학습을 사용해 LLM이 리워드 모델로부터 더 높은 점수를 받도록 학습시킨 과정을 공개

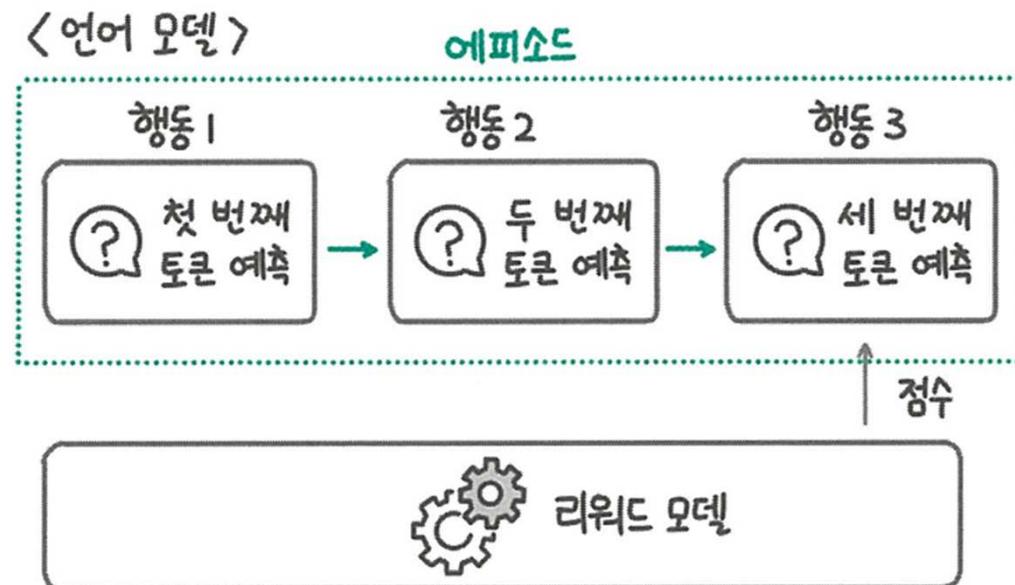


그림 4.13 강화 학습의 관점에서 언어 모델의 텍스트 평가 과정

선호도 반영하는 강화 학습

- 강화 학습

강화 학습에서는

agent가 environment에서 action을 한다.

action에 따라 environment의 state가 바뀌고, action에 대한 reward 생성 agent는 state를 인식하고 reward를 받는다.

agent는 가능하면 더 많은 reward를 받을 수 있도록 action을 수정하며 학습한다.

이 때 agent가 연속적으로 수행하는 행동의 모음을 episode 라 한다.

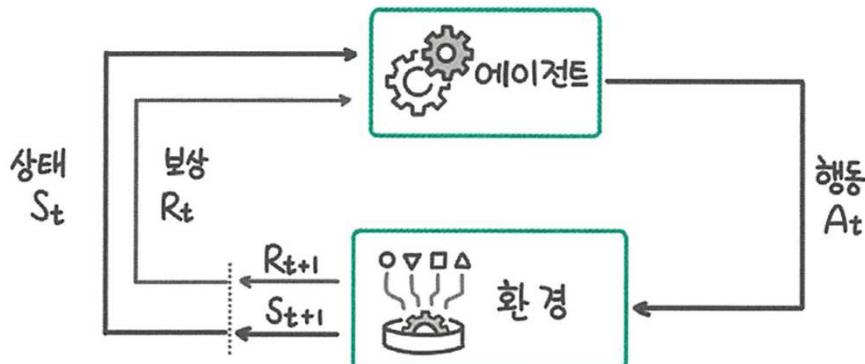


그림 4.11 강화 학습에서 에이전트가 학습하는 방식

(출처: <https://towardsdatascience.com/reinforcement-learning-101-e24b50e1d292>)

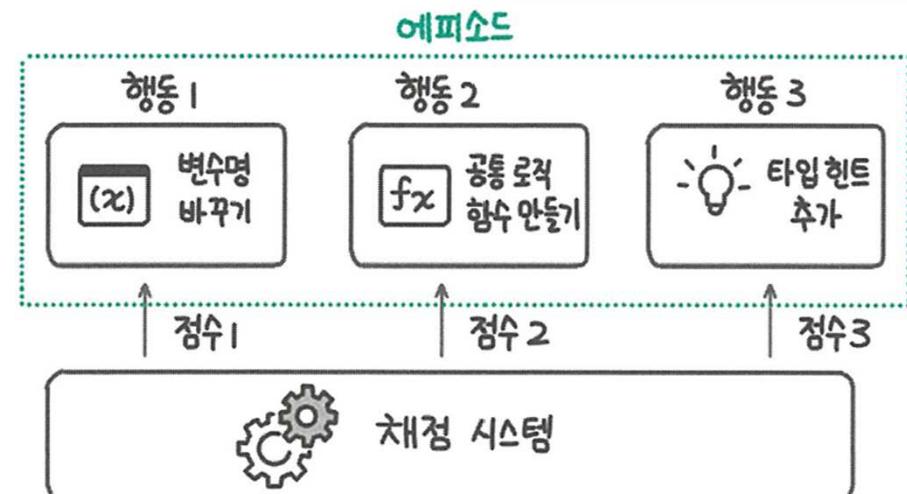


그림 4.12 강화 학습의 관점에서 나타낸 코딩 테스트 문제 풀이와 채점 과정

선호도 반영하는 강화 학습

- 강화 학습
 - 언어모델에서의 강화학습
언어 모델 : 다음 단어를 예측하는 방식으로 토큰을 하나씩 생성
강화 학습 관점)
토큰 생성 = action
언어 모델이 텍스트를 모두 생성
→ 리워드 모델이 생성한 텍스트를 생성하고 점수를 매긴다



선호도 반영하는 강화 학습

- 강화 학습
 - 언어모델에서의 강화학습
언어 모델 : 다음 단어를 예측하는 방식으로 토큰을 하나씩 생성

강화 학습 관점)

언어 모델은 행동을 취할 때마다 보상을 받지 않고,
전체 생성 결과에 대해 리워드 모델의 점수를 받는다.
이와 같은 방식으로 언어 모델은 생성한 문장의 점수가 높아지는
방향으로 학습한다.

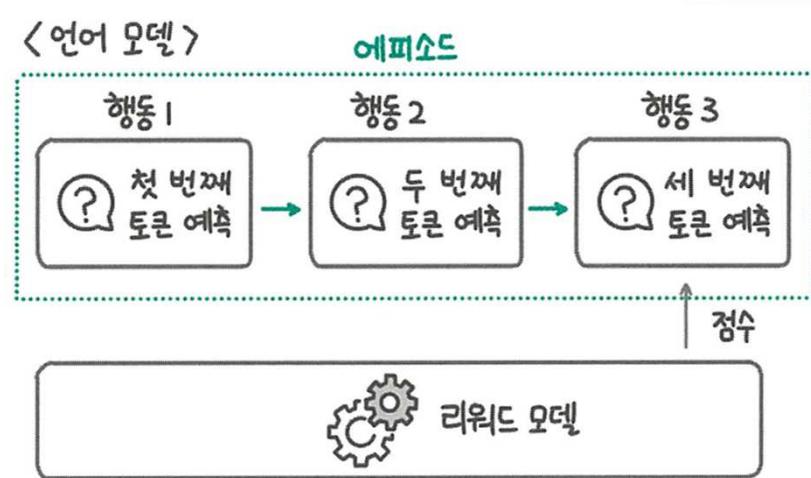


그림 4.13 강화 학습의 관점에서 언어 모델의 텍스트 평가 과정

선호도 반영하는 강화 학습

- 강화 학습 한계점_보상 해킹
 - 언어모델에서의 강화학습
언어 모델 : 다음 단어를 예측하는 방식으로 토큰을 하나씩 생성

강화 학습 관점)

언어 모델은 행동을 취할 때마다 보상을 받지 않고,
전체 생성 결과에 대해 리워드 모델의 점수를 받는다.
이와 같은 방식으로 언어 모델은 생성한 문장의 점수가 높아지는
방향으로 학습한다.

이 때, 보상을 높게 받는데에만 집중하는 보상 해킹,
reward hacking이 발생할 수 있다.

예를 들어, 코드 가독성 점수를 높게 받는 방법으로 깔끔한 코드를
작성하는 것이 아니라, 코드를 작성하지 않거나 간단한 코드만 작성해서
가독성만 높게 받는 경우가 발생할 수 있다.

선호도 반영하는 강화 학습

- PPO(Proximal Preference Optimization)

- 근접 정책 최적화 (PPO)

지도 미세 조정 모델을 기준으로 학습하는 모델이 멀지 않게 가까운 범위에서 리워드 모델의 높은 점수를 찾도록 한다.

지도 미세 조정 모델이 여기서 참고(기준) 모델이다.

EX) A : 코드 가독성이 점수가 90점, 참고 모델에서 멀어지는 경우
(reward hacking)

B : 코드 가독성이 30점, 참고 모델에서 가까운 경우

C : 코드 가독성이 80점, 참고 모델에서 가까운 경우

→ PPO는 모델 C를 찾을 수 있는 학습 방법

선호도 반영하는 강화 학습

- PPO(Proximal Preference Optimization)

- 근접 정책 최적화 (PPO)

EX) A : 코드 가독성이 점수가 90점, 참고 모델에서 멀어지는 경우

(reward hacking)

B : 코드 가독성이 30점, 참고 모델에서 가까운 경우

C : 코드 가독성이 80점, 참고 모델에서 가까운 경우

→ PPO는 모델 C를 찾을 수 있는 학습 방법

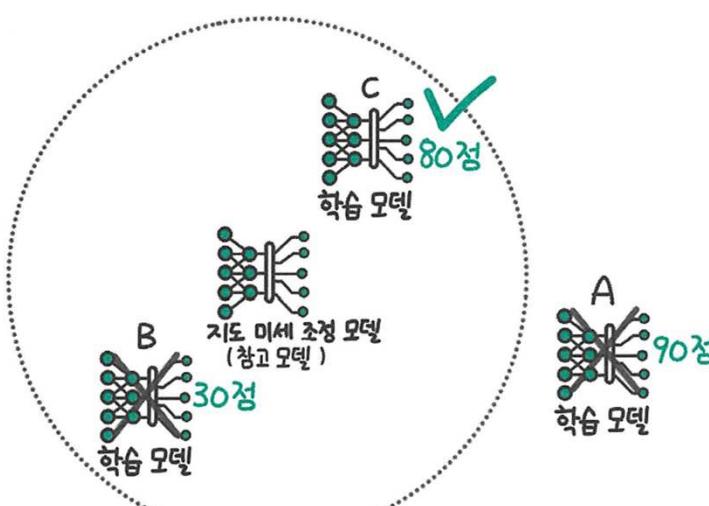


그림 4.15 참고 모델에서 멀어지지 않도록 조절하는 PPO 학습 방식

선호도 반영하는 강화 학습

- RLHF 한계점
 - 사람의 선호도를 반영한 리워드 모델을 학습한 모델, RLHF

사용법 : 리워드 모델 학습

모델 학습시킬 때 참고 모델, 학습 모델, 리워드 모델 3개 필요
-> 리소스 더 많이 필요

강화 학습 자체가 하이퍼 파라미터에 민감, 학습이 불안정

선호도 반영하는 비강화 학습

- 강화 학습이 꼭 필요할까?

- 기각 샘플링 :

여러 답변 생성 결과 중 리워드 모델이 가장 높은 점수를 준 결과 중 리워드 모델이 가장 높은 점수를 준 결과를 LLM의 지도 미세 조정에 사용하는 방법

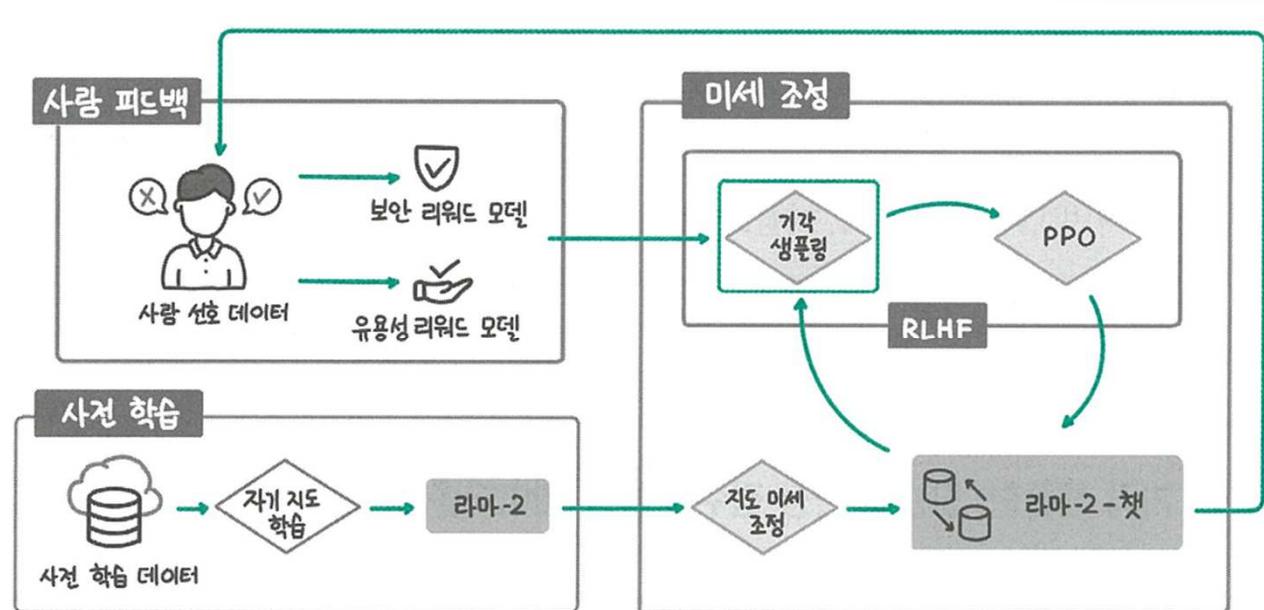


그림 4.16 라마-2의 학습 과정에 사용된 기각 샘플링(출처: <https://arxiv.org/pdf/2307.09288.pdf>)

선호도 반영하는 비강화 학습

- 강화 학습이 꼭 필요할까?
 - DPO (Direct Preference Optimization) :
 DPO에서는 선호 데이터 셋을 직접 언어 모델에 학습시킨다.

리워드 모델 필요X

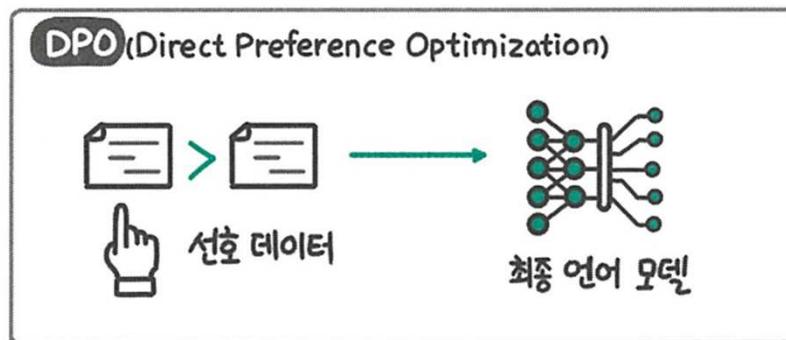
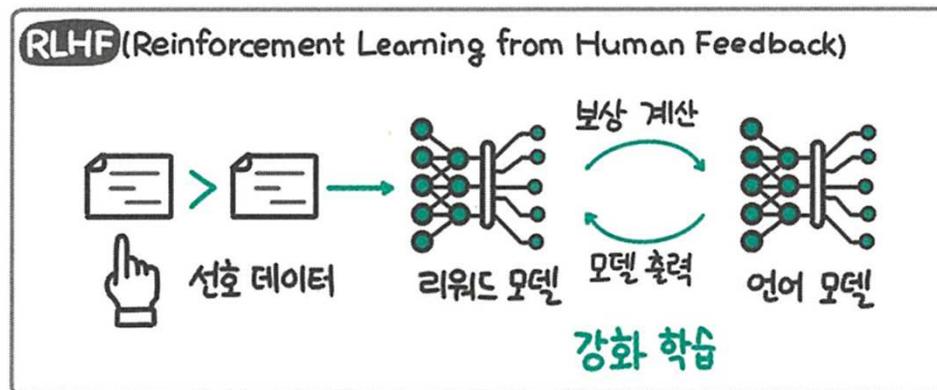


그림 4.17 DPO와 RLHF의 차이(출처: DPO 논문, <https://arxiv.org/pdf/2305.18290.pdf>)

정리

- 정리

- 언어모델 :

사전 학습 -> 지시 데이터셋으로 지도 미세 조정 -> 지시 사항에 응답이 원활
-> 사람이 선호하는 방식으로 답변을 생성 (선호 데이터 셋을 활용)

선호 학습 방법 :

강화 학습, 비강화학습

강화 학습-하이퍼파라미터에 민감, 학습에 불안정

비강화학습-기각 샘플링, DPO



