4. 말 잘 듣는 모델 만들기

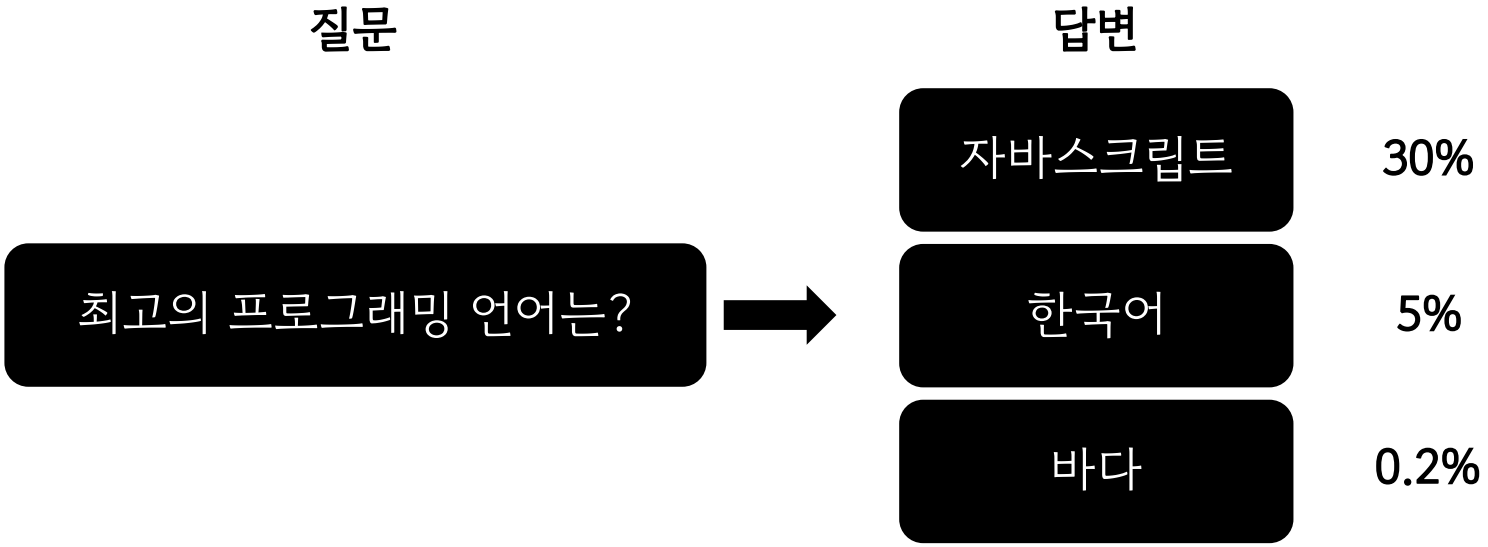
김근영

사용자의 말에 이어질 법한
텍스트를 생성

사용자의 요청에 맞게
응답

1. 사용자의 요청에 응답할 수 있도록 학습 : 사전학습, 지도 미세 조정
2. 사용자가 더 선호하는 답변 생성하도록 모델 조정
- 강화학습을 사용하는 방법 : RLHF, PPO
- 강화학습을 사용하지 않는 방법 : 기각 샘플링, 직접 선호 최적화

1.1 사전학습



1.2 지도 미세 조정

데이터

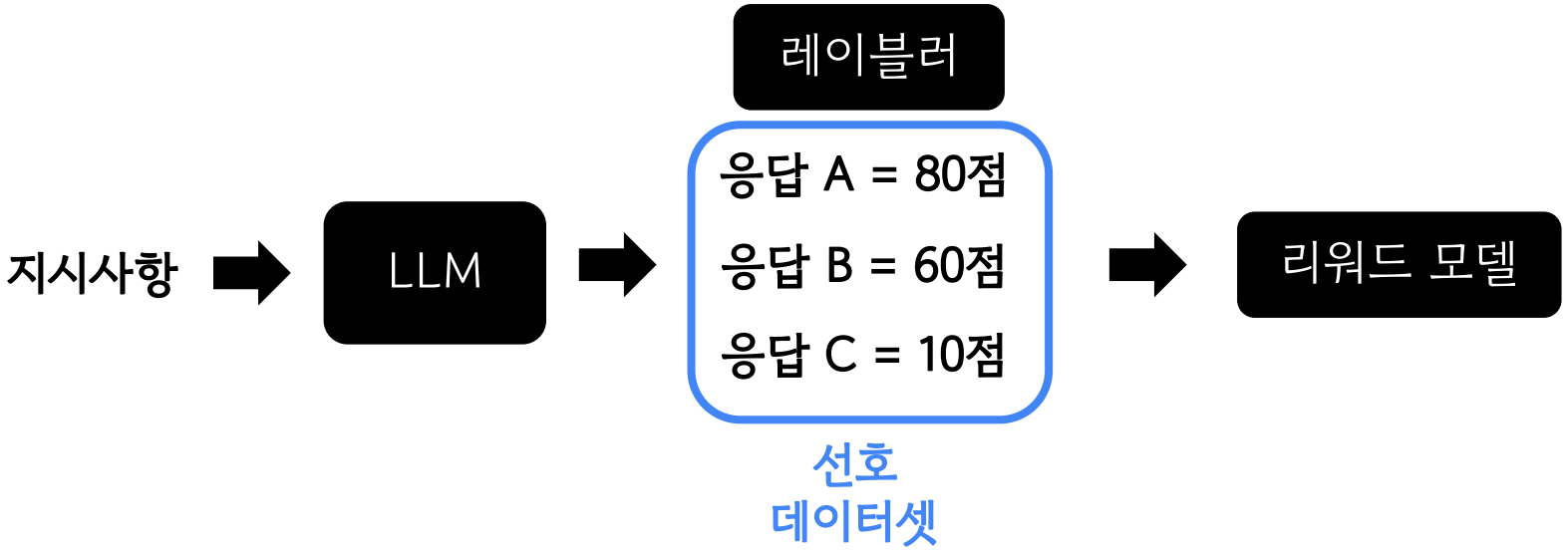

```
{
  "instruction": "Create a classification task by clustering the given list of items.",
  "input": "Apples, oranges, bananas, strawberries, pineapples",
  "output": "Class 1: Apples, Oranges\nClass 2: Bananas, Strawberries\nClass 3: Pineapples",
  "text" : "Below is an instruction that describes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


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}
"""
```

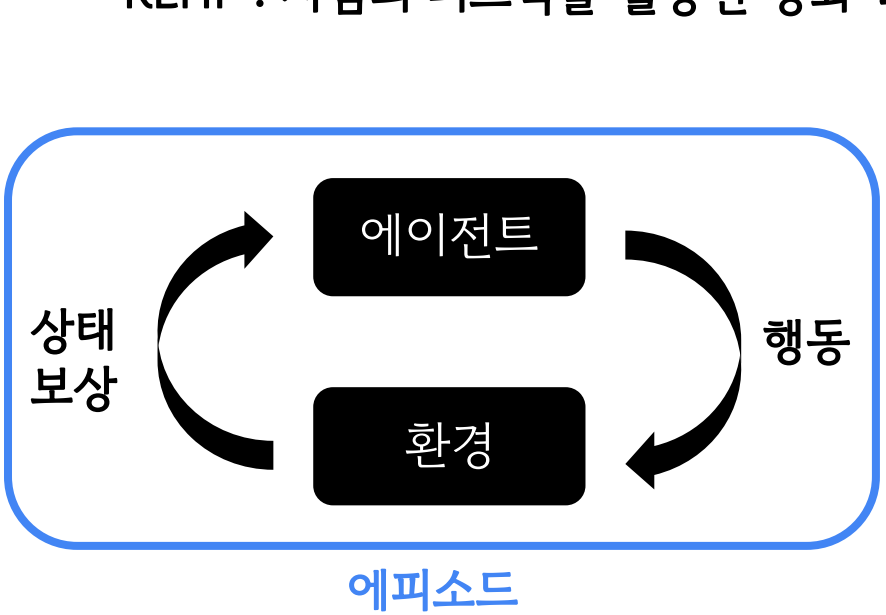
→ **인과적 언어 모델링**

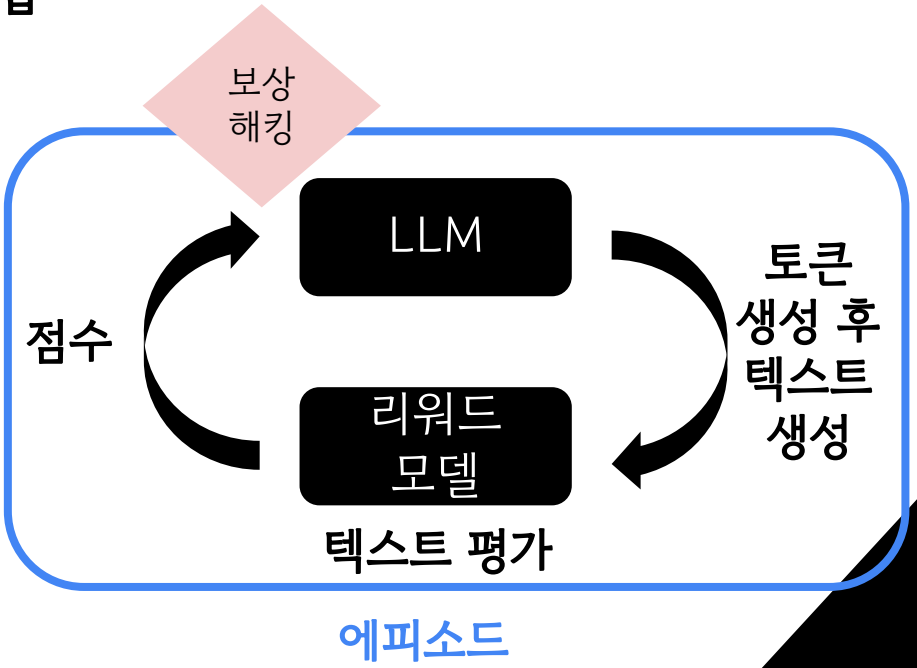
2.1 선호 데이터셋



2.2 강화 학습

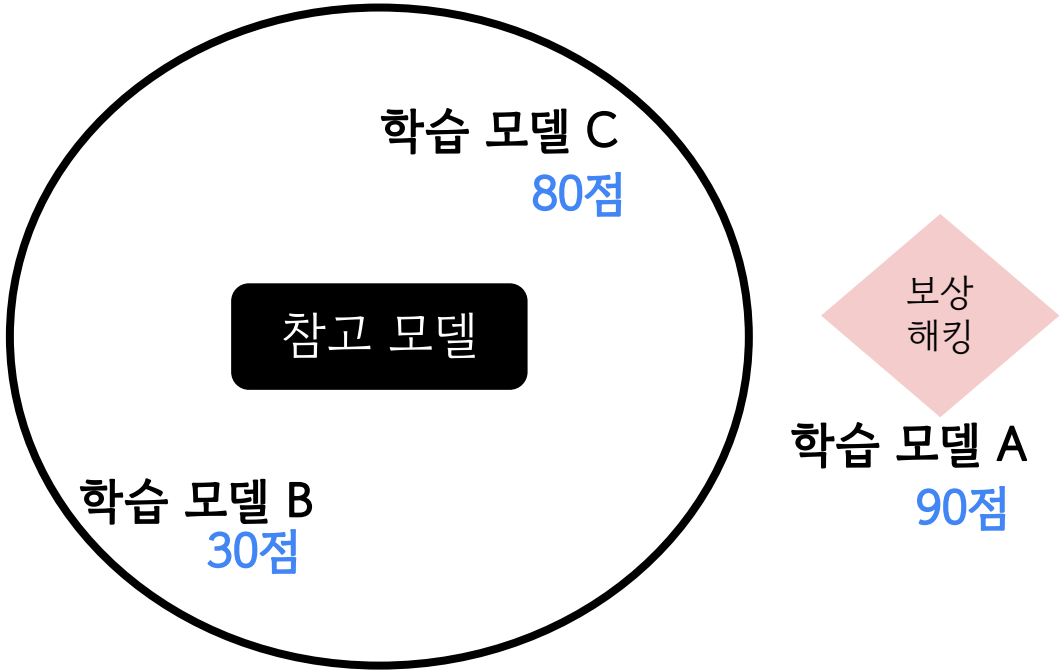
RLHF : 사람의 피드백을 활용한 강화 학습



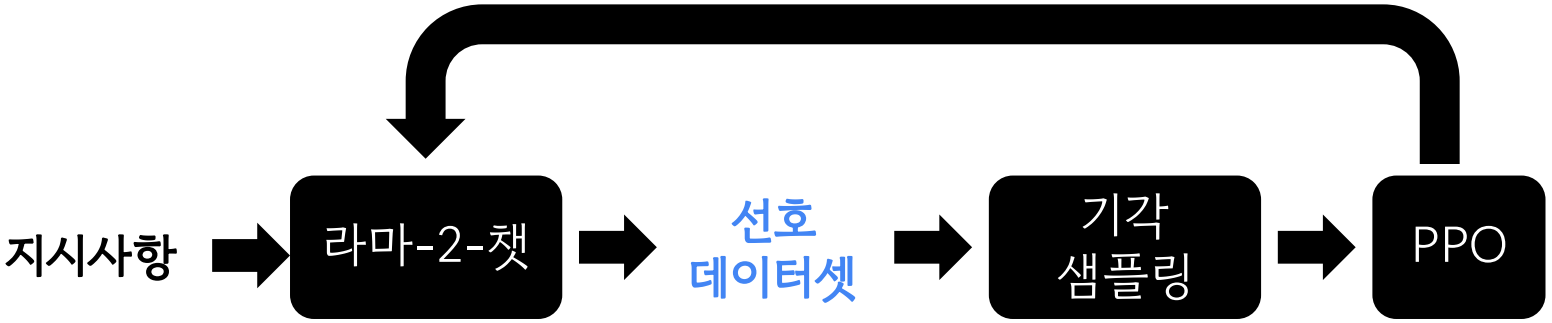


2.3 PPO

PPO(Proximal Preference Optimization) : 근접 정책 최적화



3.1 기각 샘플링



3.2 DPO(직접 선호 최적화)

최고의 프로그래밍 언어는?

자바스크립트	30%	선호
한국어	5%	비선호
바다	0.2%	비선호

자바스크립트	33%
한국어	2%
바다	0.1%

3.3 DPO를 사용한 모델

Model Card for Zephyr 7B β

Zephyr is a series of language models that are trained to act as helpful assistants. Zephyr-7B-β is the second model in the series, and is a fine-tuned version of mistralai/Mistral-7B-v0.1 that was trained on on a mix of publicly available, synthetic datasets using Direct Preference Optimization (DPO). We found that removing the in-

허깅페이스 팀의 **제퍼-7B-베타**

- 4개의 LLM이 생성한 결과를 **AI가 평가한 뒤** 선호/비선호 데이터쌍 구축
- 사람의 평가가 아닌 AI 평가로도 잘 동작한다는 사실이 확인됨