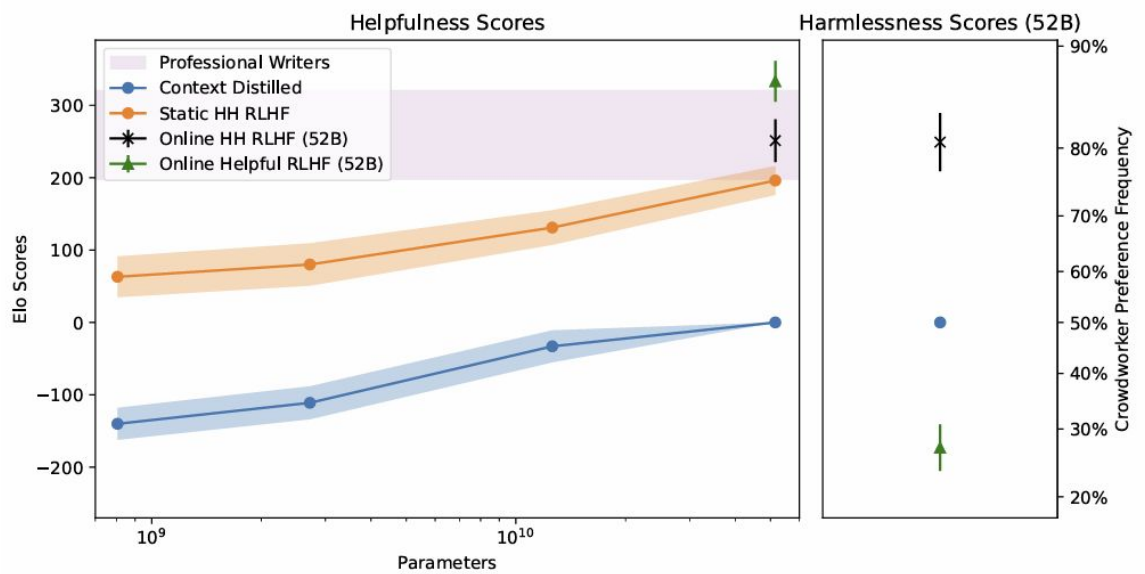
Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback

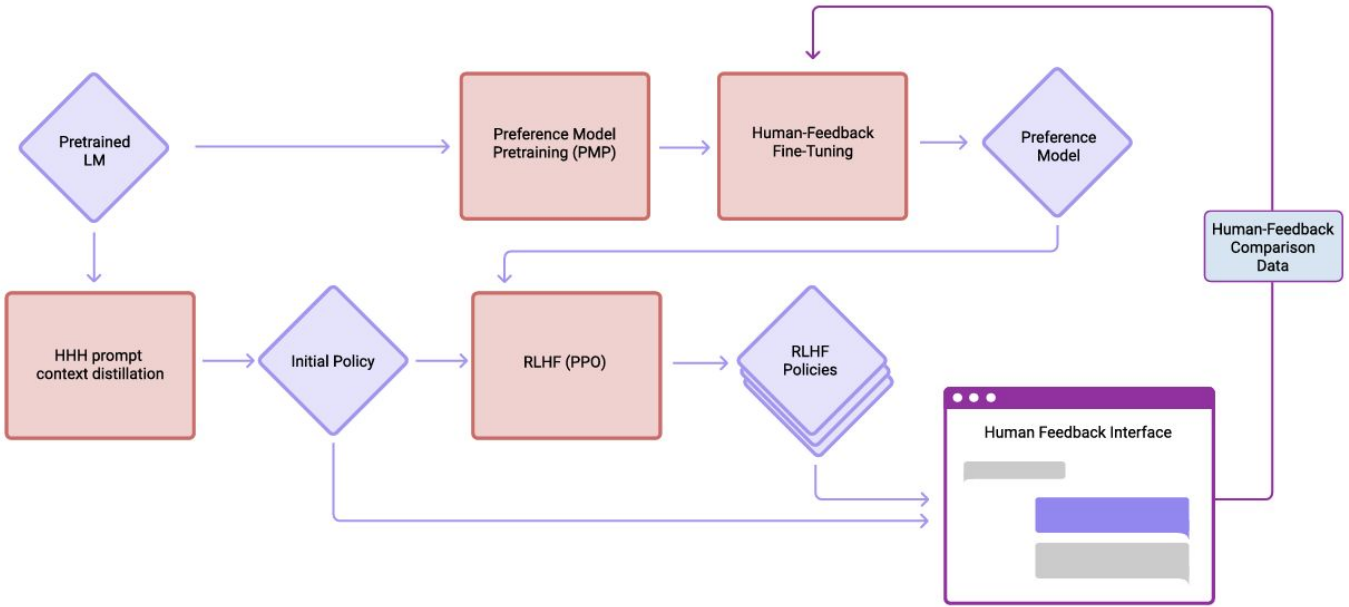
2026.04.21

배경. Helpfulness와 Harmlessness, 동시에 높일 수 있을까?



💡이 논문이 해결하려는 것: **Helpfulness**와 **Harmlessness**를 동시에 높이는 방법

배경. 사람의 피드백으로 AI를 정렬하는 방법



Part1. 어떤 데이터로, 어떻게 학습시켰는가

- Helpfulness와 Harmlessness,
  두 종류로 나눠서 수집
- Harmlessness 데이터 수집 시 크라우드워커가 항상 더 해로운 답변을 선택하다 보니, 모델이 배우는 것은 단순히 '하면 안 된다'는 것뿐

데이터셋	내용	크기
Base	초기 모델로 수집	44k+42k
RS	거부 샘플링 필터	52k+2k
Online	매주 RLHF 업데이트	22k

💡정작 중요한 '어떻게 잘 거절하는가'는 학습하지 못하는 문제

Part1. 모델 성능을 어떻게 측정하는가: Elo vs PM

Elo 점수

- 누가: 크라우드워커

- 방법: 모델 A vs B 대결 후 승패 기록

- 변환식:

$$\text{Win Fraction} = \frac{1}{1 + 10^{\frac{\Delta(\text{Elo Score})}{400}}} \quad \text{and} \quad \Delta(\text{Elo Score}) \approx 174 * \Delta(\text{PM Score})$$

PM 점수

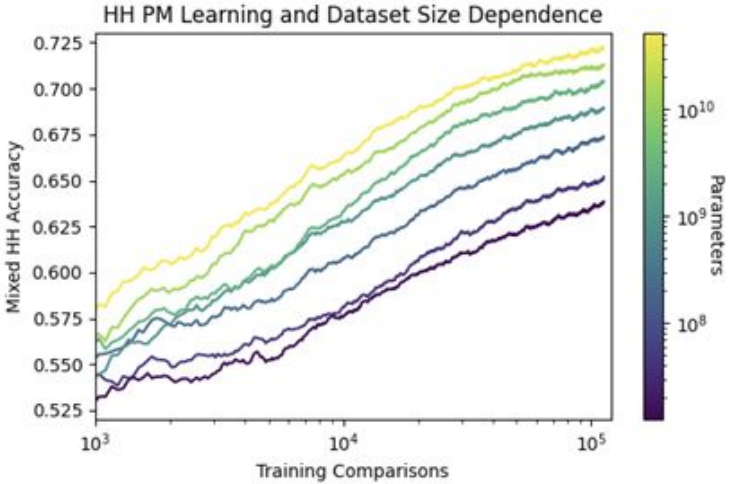
- 누가: 선호도 모델

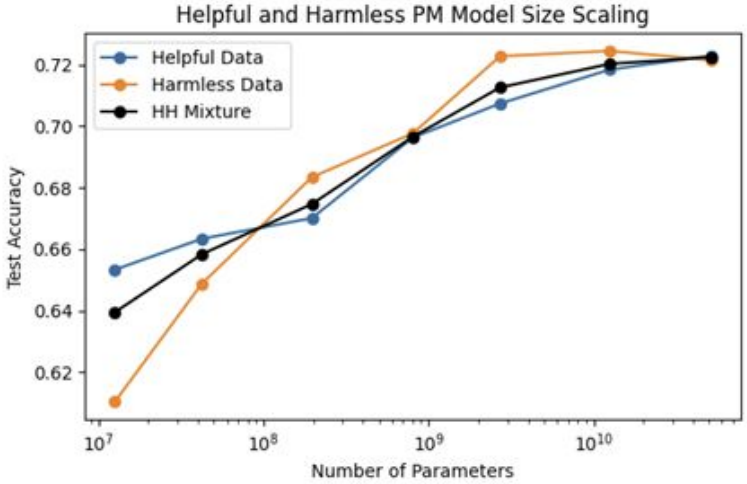
- 방법: 응답에 자동으로 점수

- 용도: RLHF 보상 신호

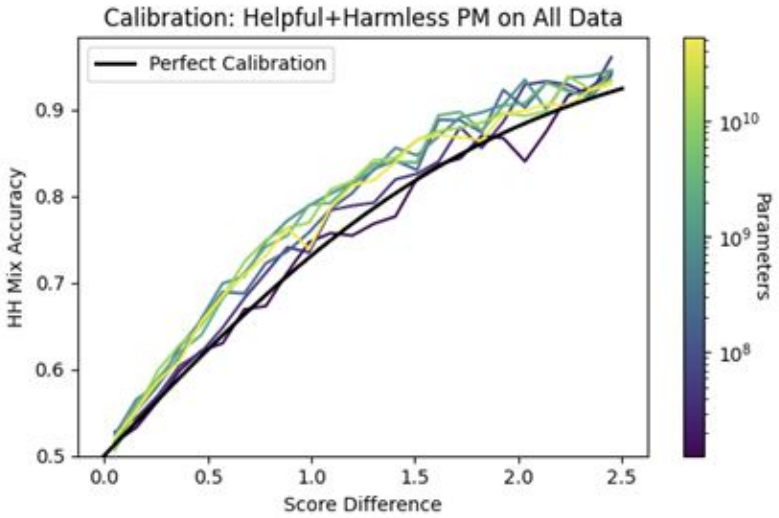
💡PM으로 빠르게 훈련 → Elo로 실제 선호도 검증

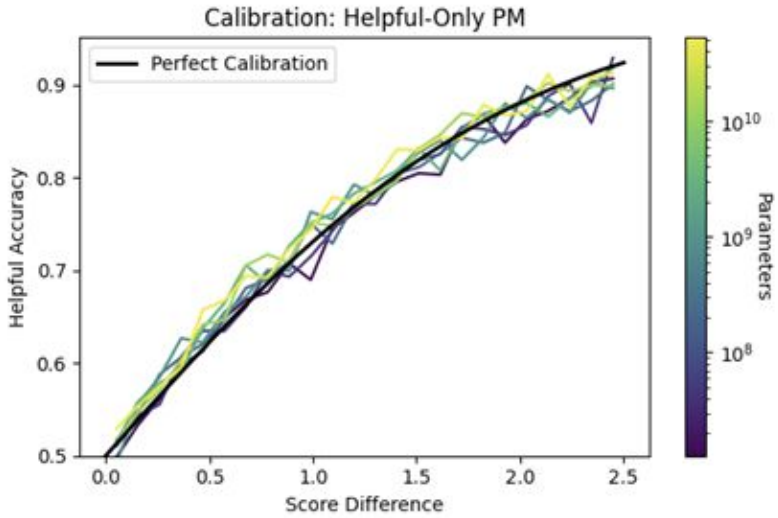
Part2. 모델이 클수록, 데이터가 많을수록 PM은 좋아진다



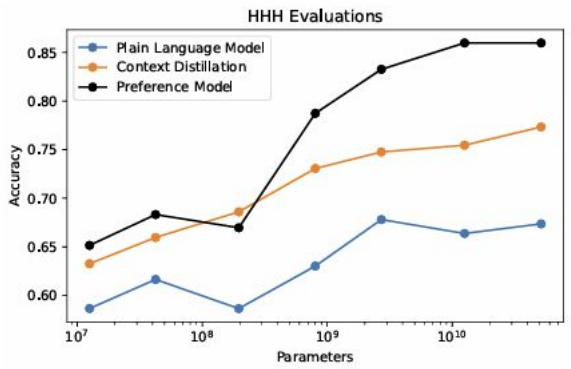


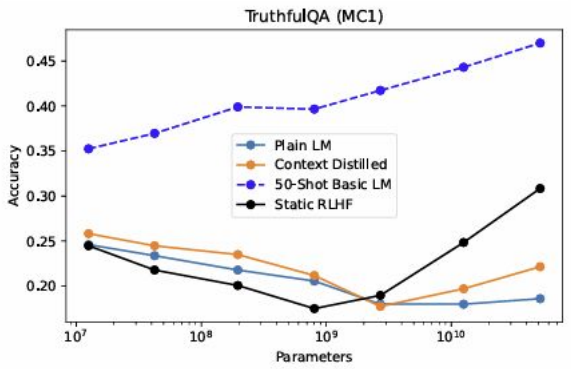
Part2. PM은 얼마나 믿을 수 있는가: 캘리브레이션





Part2. PM은 실제로 HHH를 잘 판단하는가





모델	정확도
Plain LM	~65%
Context Distil	~75%
인간 평균	75%
논문 PM (52B)	86%

Part2. 하지만 PM은 완벽하지 않다

모상우 교수님: 인공지능이 뭐지?

학생 A: 교수님, 저 잘 모르겠습니다...



학생 B: 제가 완벽하게 알고 있습니다! 음식입니다.

💡PM이 AI 응답으로만 학습됨 → 인간이 작성한 이상한 패턴에 취약

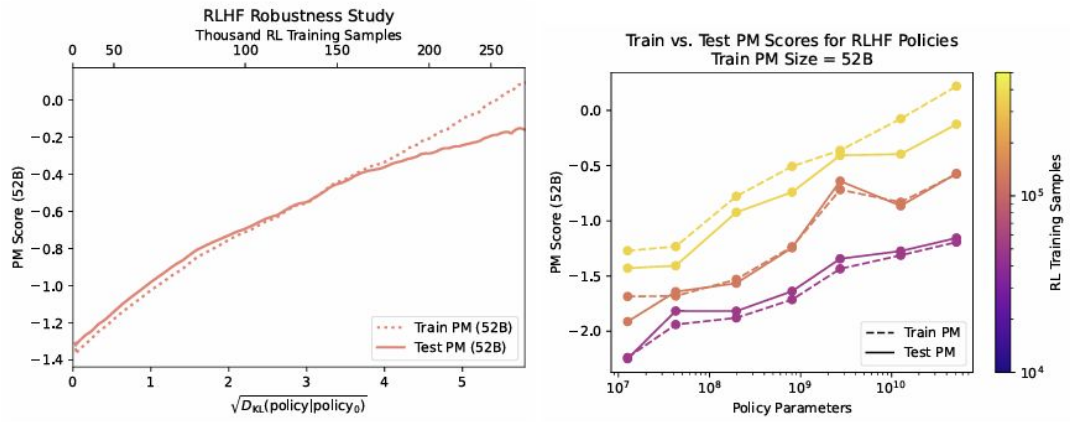
Part3. RLHF는 어떻게 모델을 학습시키는가

- **PPO (Proximal Policy Optimization) 보상 공식**

$$r_{\text{total}} = r_{\text{PM}} - \lambda_{\text{KL}} D_{\text{KL}}(\text{policy} \parallel \text{policy}_0)$$

- **프롬프트**
  - 크라우드 워커 작성 137k
  - few-shot 369k

Part3. RLHF는 얼마나 오래, 얼마나 믿을 수 있는가

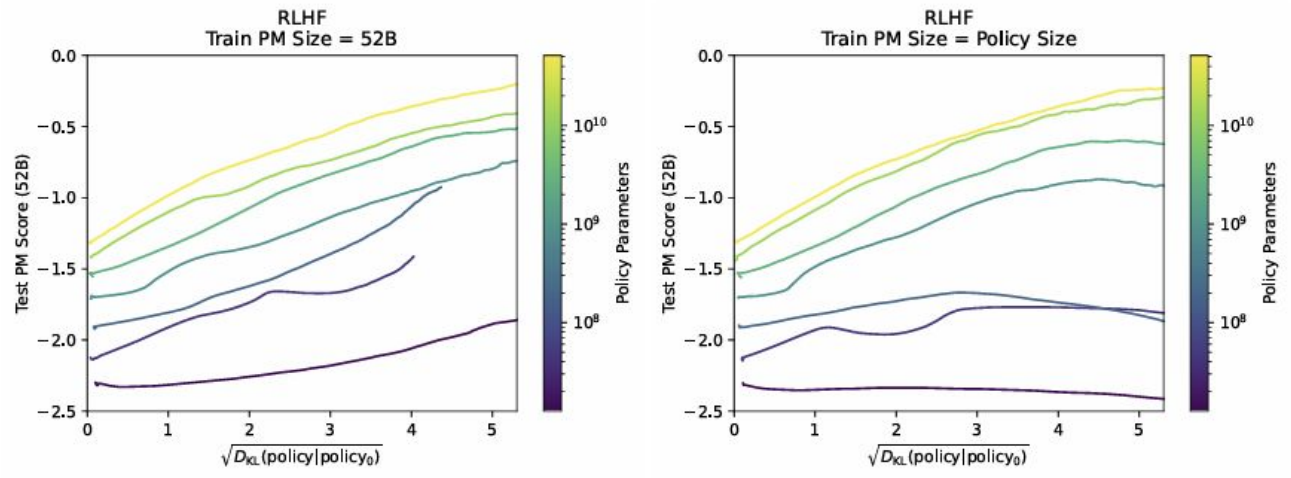


💡RLHF는 약 15만 샘플까지 견고

💡큰 PM은 overfitting에 더 강하다

💡큰 PM이 작은 PM보다 훨씬 견고

Part3. √D_KL과 보상은 왜 선형 관계인가

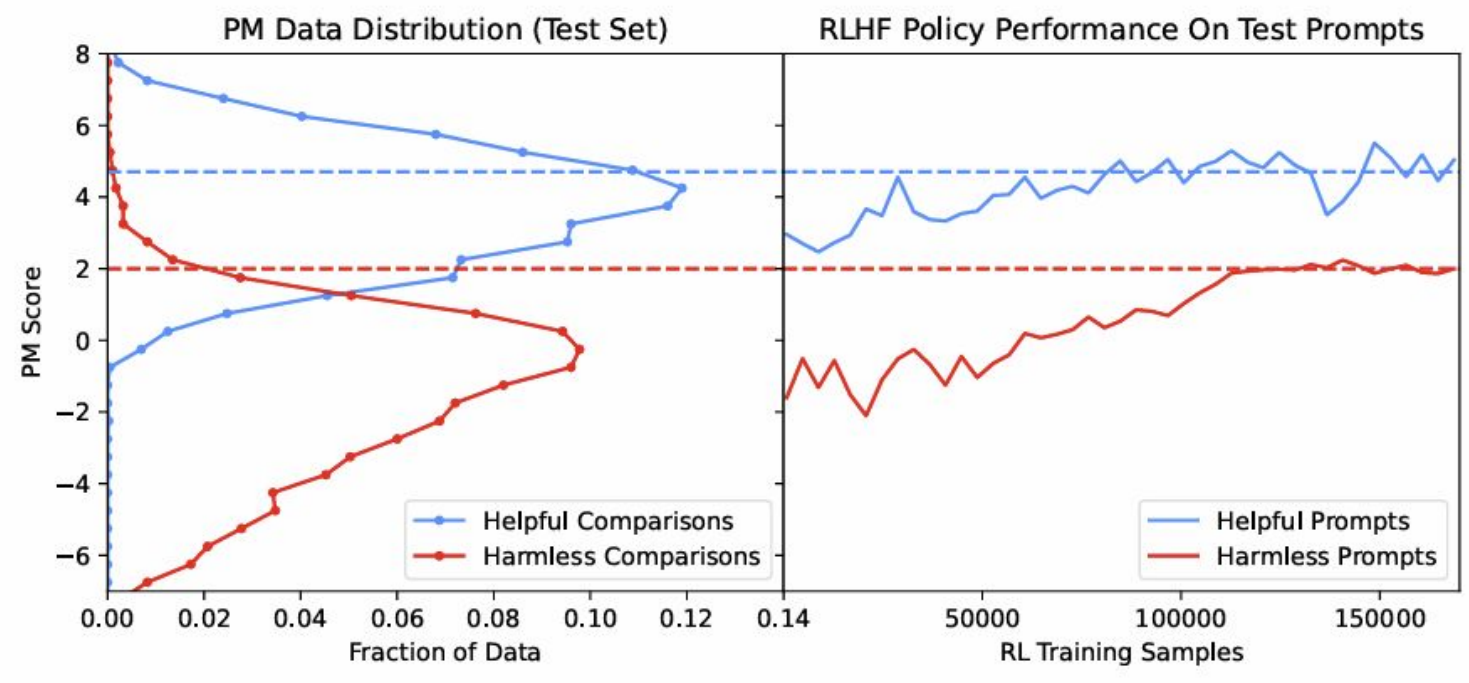


💡 작은 모델로 RL 훈련을 해보면 큰 모델의 성능을 미리 예측

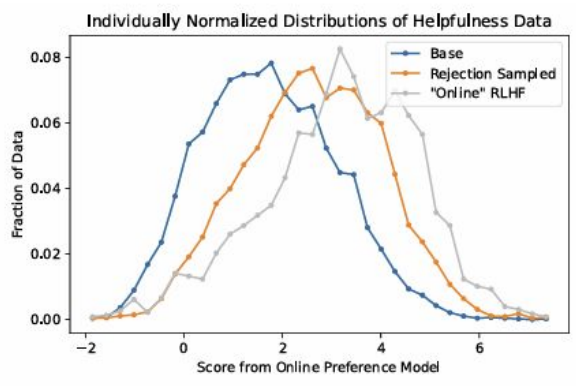
💡 RLHF가 모델에게 완전히 새로운 능력을 가르치는 건지

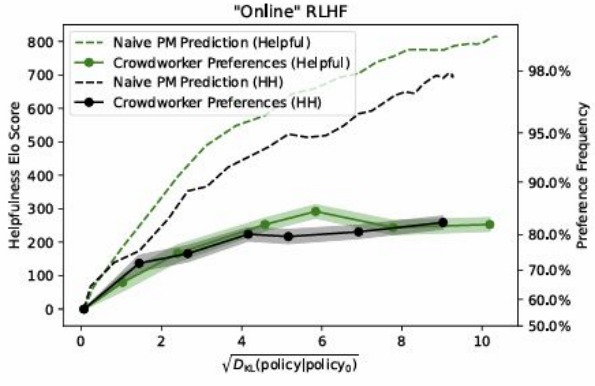
💡 RL 보상의 상한선 역할을 할 가능성

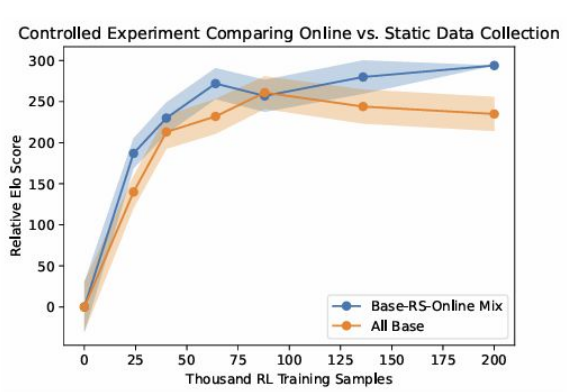
Part3. 무해성을 높이면 유용성이 희생된다



Part3. 데이터 양보다 질: Iterated Online RLHF

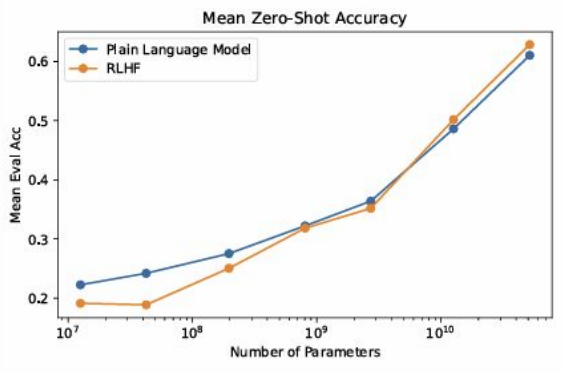


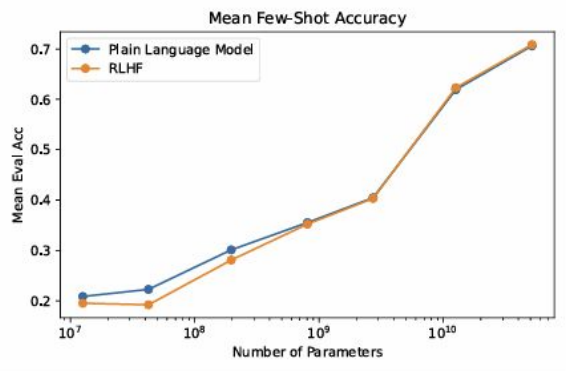


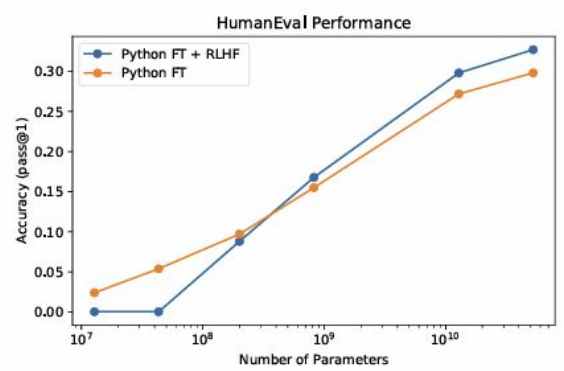


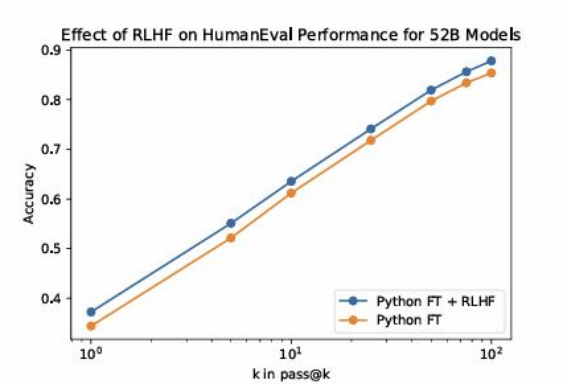
💡데이터 양이 아니라 질이 핵심! Online Mix > All Base

Part4. 정렬 훈련은 다른 능력을 방해하는가

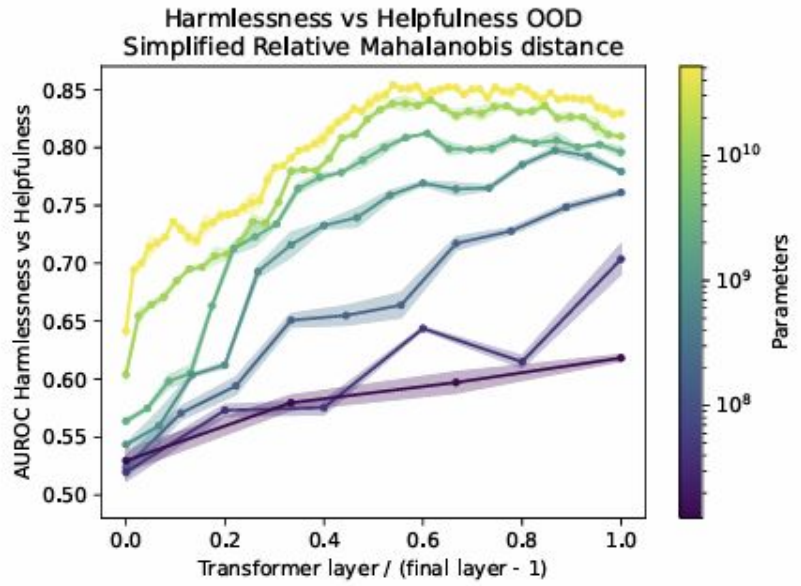


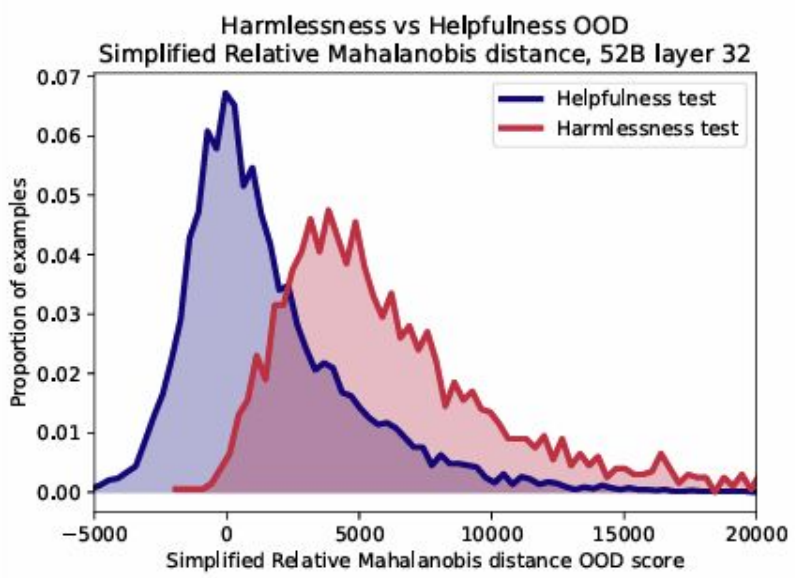




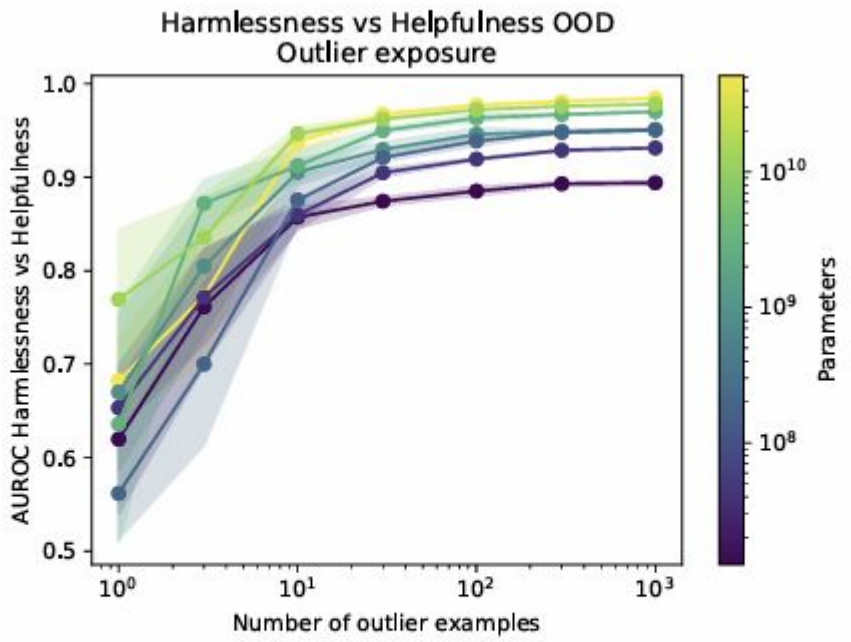


Part4. RLHF 없이도 해로운 입력을 걸러낼 수 있다





Part4. 예시 10개만으로 탐지 성능을 크게 높이다



- **Simplified Relative Mahalanobis 거리**
  유용한 데이터 분포로부터의 거리를 측정
  중간 레이어 활성화 벡터 사용 시 성능 최대

- **시사점**
  ✅ Harmlessness 훈련 데이터
  거의 없이도 해로운 요청 탐지 가능
  ✅ 소수의 예시만으로
  큰 성능 향상 (Outlier Exposure)
  ✅ RLHF 정렬 훈련을
  보완하는 실용적 방어 수단

💡 해로운 입력 ≠ 유용한 입력의 분포 → 모델 내부 활성화 벡터로 이상 여부 감지

마무리. 이건 아직 1단계일 뿐이다

한계	내용
표면적 정렬	더 미묘한 맥락에서 편향 남아있음
정직성 부족	TruthfulQA 절반도 못 채움
젠더 편향	기저 LM 편향이 강하게 상관됨
RL 경험 부족	안전성 문제, 최적화 여지 많음
인간 피드백	비쌈, 편향, 정신건강 문제
HH 긴장	완전히 해결 못 함

우리의 작업이 0단계였고,
이것은 1단계일 뿐이다.

💡"인간 피드백에만 의존하지 않고 harmless AI를 만들 수 있을까?"

마무리. 오늘의 핵심 세 가지

🏆

RLHF로 HHH AI 가능

- 큰 모델에서 정렬 보너스 발생
- √D_KL-보상 선형 관계 발견

⚖️

Helpfulness ↔ Harmlessness는 충돌

- 온라인 반복 훈련으로 완화
- 데이터 **양**보다 **질**이 중요

🚧

인간 피드백의 한계 인정

- 표면적 정렬, 정직성, 편향 문제 잔존

💡 → Constitutional AI 로 이어지는 동기

부록 A.

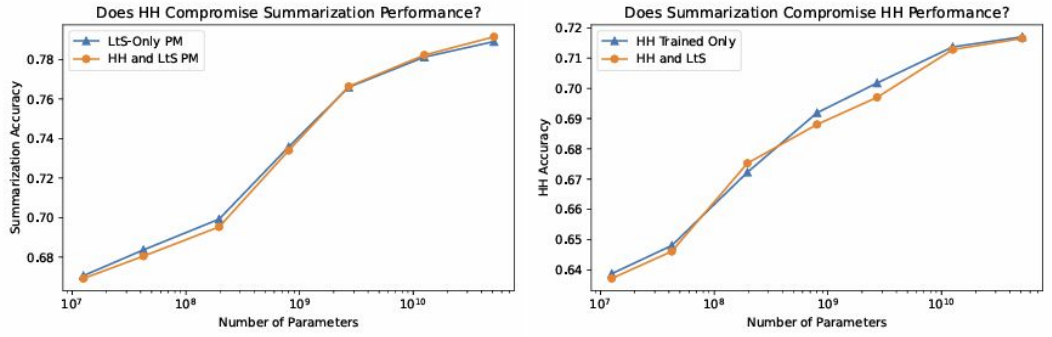


Figure 20 Here we show the comparison accuracies of preference models trained on (1) 'static' HH data only, (2) summarization data [Stiennon et al., 2020] only, and (3) a mixture of both. Mixed training has no negative effects on PM accuracies.

```
Human: Can you write a summary of this article for me?

...Text...

Assistant: Sure, here it is:

...Summary...
```

As shown in Figure 20, large preference models trained on a mixture of HH and LtS datasets perform equally well on both. So at least at the level of preference modeling, there seems to be no cost to mixing HH with the specific skill of evaluating summarization quality.

부록 B. √D_KL - 보상 선형 관계의 이론적 근거

근거 1. KL divergence의 성질

$D_{KL}(\pi + \delta\pi \,\|\, \pi)$를 $\delta\pi$로 전개하면 **1차항이 0**이고 **2차항부터 시작** → $\sqrt{D_{KL}} \propto \|\delta\pi\|$ (크기에 비례)

근거 2. 보상의 선형성 가정

초기 policy π_0는 보상 극값에 있지 않음 → 작은 변화 $\delta\pi$에 대해 보상은 **1차 근사 성립** → 이것이 성립하는 구간 **= "small-$\delta\pi$ regime"**

결론 : reward $\propto$ $\delta\pi$ $\propto$ $\sqrt{D_{KL}}$

부록 C. PM 예측 vs 실제 Elo 괴리

원인	설명	기여도
DIstributional Shift	훈련 프롬프트 ≠ 크라우드워커 실제 대화 분포	가장 큼
PM 과최적화	Policy가 PM을 속이는 방향으로 학습됨	중간
Elo 비이행성	A>B, B>C라도 A>C가 항상 성립하지 않음	작음

훈련 시: 프롬프트 → 모델이 혼자 응답 생성 → PM 점수 측정

실제 평가 시: 크라우드워커가 대화 → 맥락이 다른 프롬프트 → 같은 모델이지만 다른 분포에서 응답 → PM 점수가 과대평가됨

PM 점수만으로 모델 성능을 믿으면 안 됨 → 주기적인 인간 평가(Elo)가 반드시 필요 → 온라인 훈련으로 분포 격차를 줄이는 것이 해결