Learning to summarize from human feedback (2020)

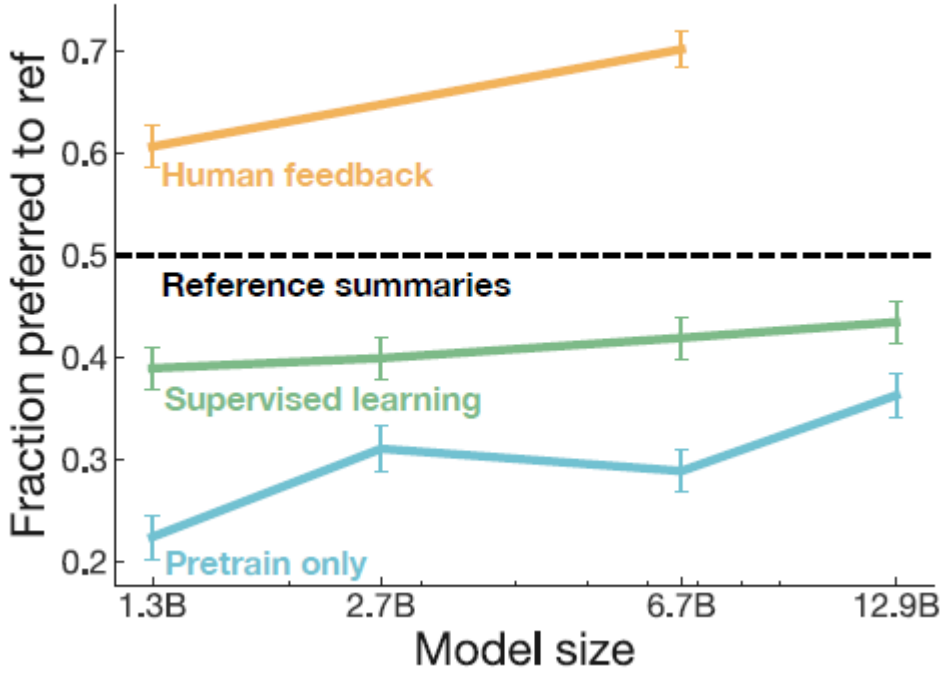
Stiennon, Nisan, et al. Advances in Neural Information Processing Systems 33 (2020): 3008-3021. Open AI

IIE8557-01 고등강화학습

경영과학연구실 이태헌

As language models become more powerful, training and evaluation are increasingly bottlenecked

- 언어 모델이 커지면서, 학습과 평가에 병목 현상을 겪고 있음
- 모델이 인간의 선호도를 최적화하도록 학습함으로써 이러한 문제를 개선시킬 수 있음



‘Human feedback을 반영하여, 더 높은 품질의 요약을 생성하는 언어 모델을 만들고자 함’

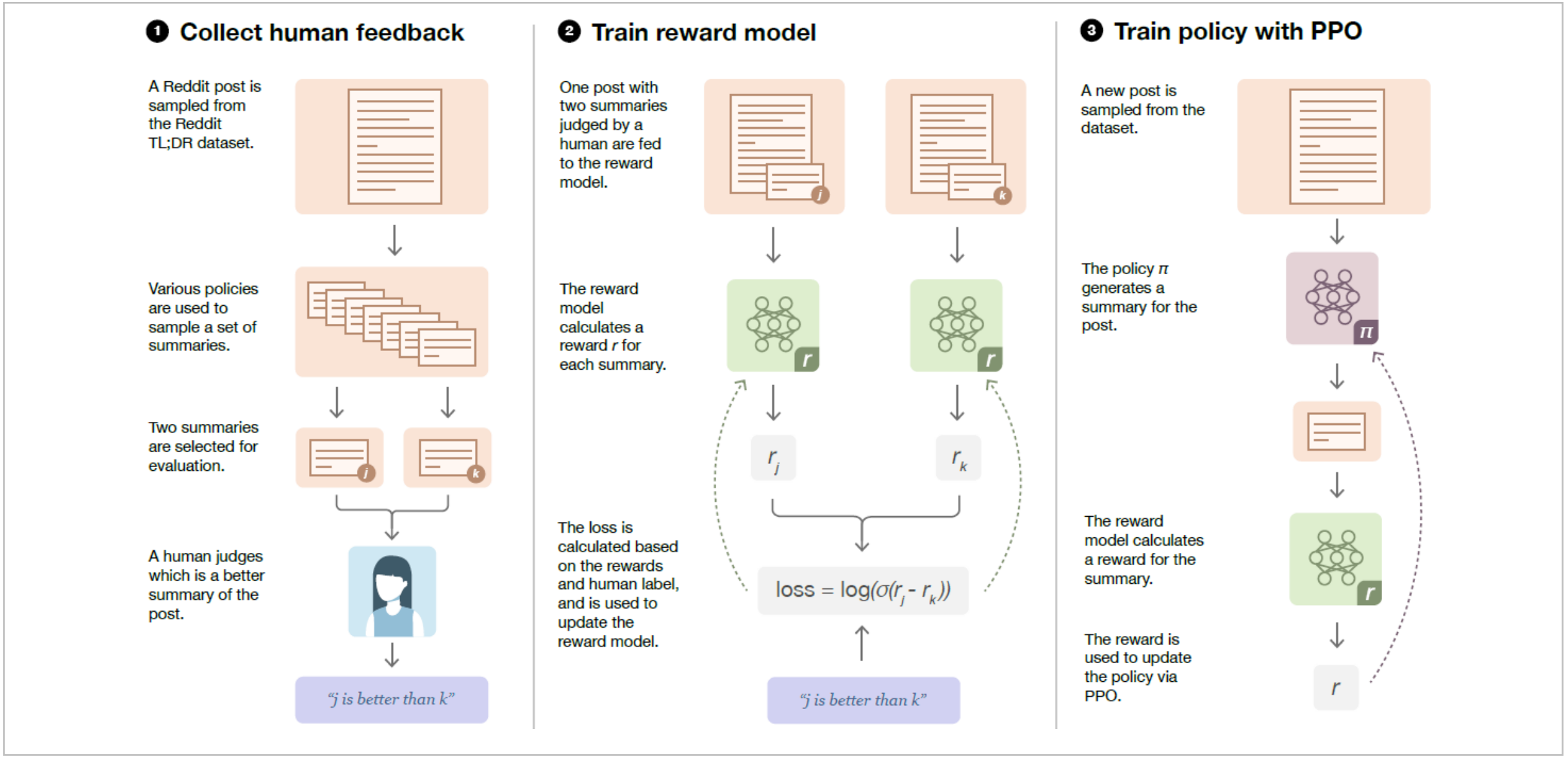
Human feedback 데이터 수집: 두 가지 요약 사이의 선호도를 비교하는 Human feedback에 기반한 대규모 데이터셋 수집

보상 모델 학습: 수집된 Human feedback 이용해, 어떤 요약이 더 선호되는지를 예측하는 보상 모델 학습
이 모델은 각 요약이 주어진 기준에 얼마나 잘 부합하는지에 대한 점수 출력함

강화 학습을 이용한 정책 훈련 : 학습된 보상 모델 사용하여 요약을 생성하는 정책(policy)을 강화 학습 통해 Fine-tuning.
이 정책은 각 시간 단계마다 텍스트 토큰을 생성하며, 보상 모델로부터 주어진 전체 생성된 요약에 대한 '보상'에 기반하여 업데이트 됨

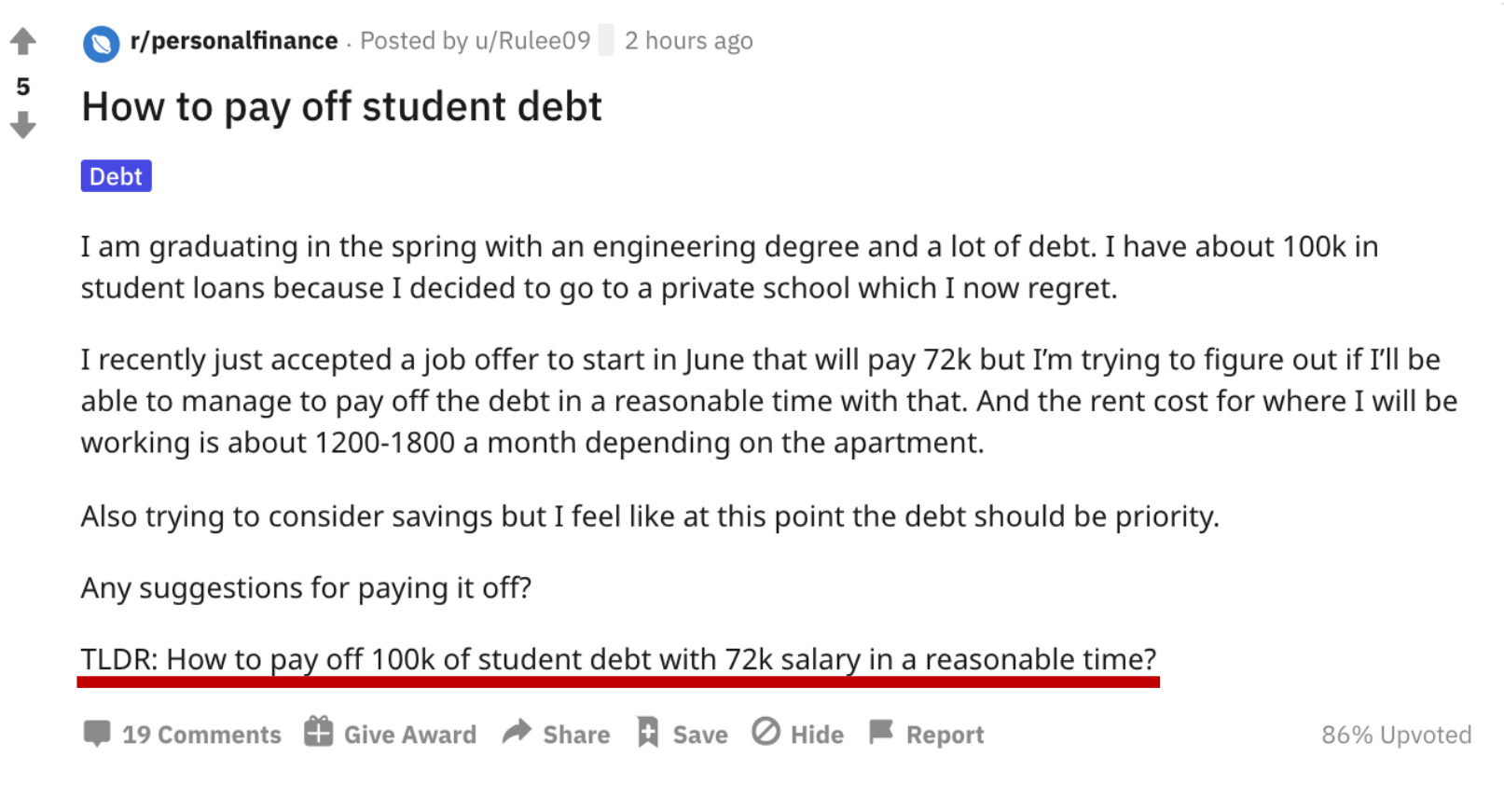
Human feedback 통한 연속적 개선: 강화 학습을 통해 얻은 요약 결과를 사용하여 추가적인 Human feedback 을 수집하고, 이 과정을 반복하여 모델을 지속적으로 개선

모델 프로세스


1 Collect human feedback
A Reddit post is sampled from the Reddit TL;DR dataset.
Various policies are used to sample a set of summaries.
Two summaries are selected for evaluation.
A human judges which is a better summary of the post.
"j is better than k"
2 Train reward model
One post with two summaries judged by a human are fed to the reward model.
The reward model calculates a reward r for each summary.
The loss is calculated based on the rewards and human label, and is used to update the reward model.
loss = log(σ(r_j - r_k))
"j is better than k"
3 Train policy with PPO
A new post is sampled from the dataset.
The policy π generates a summary for the post.
The reward model calculates a reward for the summary.
The reward is used to update the policy via PPO.

TL;DR Dataset (Too Long; Didn't Read)

- **Reddit.com의 다양한 주제에 걸쳐 있는 약 300만 개의 게시물과 원글 작성자가 작성한 게시물의 요약본(TL;DR) 포함**



r/personalfinance · Posted by u/Rulee09 2 hours ago

5

How to pay off student debt

Debt

I am graduating in the spring with an engineering degree and a lot of debt. I have about 100k in student loans because I decided to go to a private school which I now regret.

I recently just accepted a job offer to start in June that will pay 72k but I'm trying to figure out if I'll be able to manage to pay off the debt in a reasonable time with that. And the rent cost for where I will be working is about 1200-1800 a month depending on the apartment.

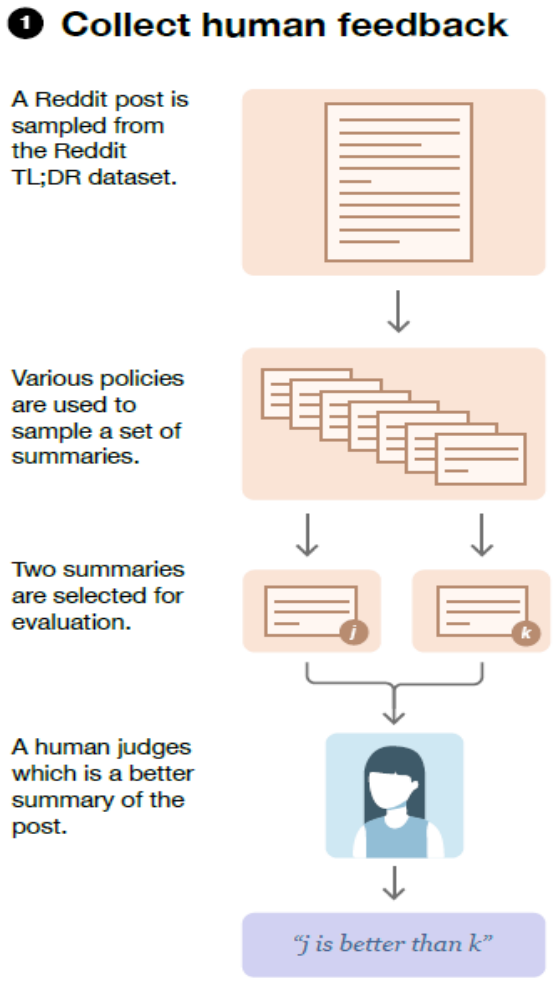
Also trying to consider savings but I feel like at this point the debt should be priority.

Any suggestions for paying it off?

TLDR: How to pay off 100k of student debt with 72k salary in a reasonable time?

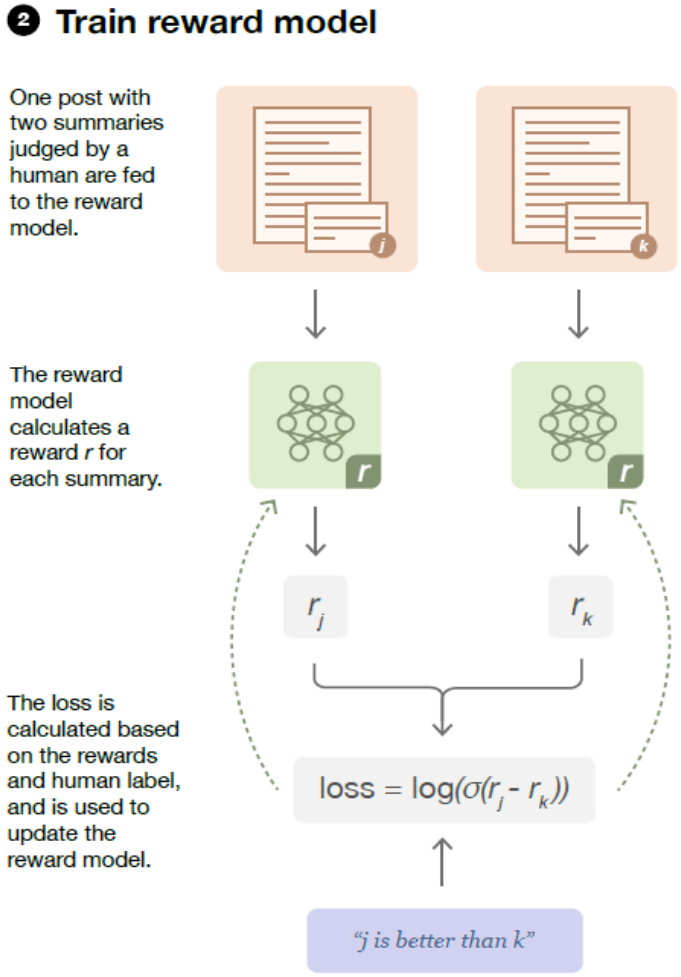
19 Comments Give Award Share Save Hide Report 86% Upvoted

Step 1 : collect samples from existing policies and send comparisons to humans



- 각 Reddit post에 대해 Current Policy, Initial Policy, Original reference summaries, Various baselines 포함한 여러 방법으로 summary sampling
- Summary pair가 human evaluators에게 보내지며, 그들은 주어진 reddit post의 best summary 선택
- 이 과정은 Offline learning으로 진행됨

Step 2 : Learn a reward model from human comparisons

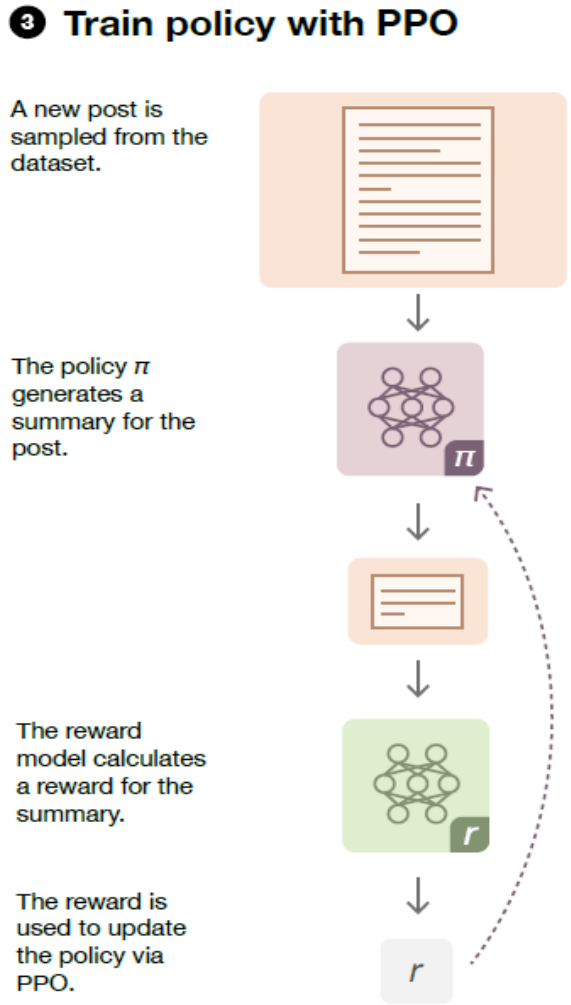


- Post와 후보 summaries가 주어지면, reward model이 summary가 human evaluator의해 더 낫다고 판단될 로그 확률을 예측하도록 학습됨
- 모델은 post x가 주어졌을 때, summary y $\in \{y_0, y_1\}$ 중 어느 것이 사람에 의해 더 낫다고 평가되는지 예측
- Human evaluator가 선호하는 summary가 y_i인 경우, reward model의 loss function은 아래 식과 같음

$$\text{loss}(r_\theta) = E_{(x,y_0,y_1,i)\sim D}[\log(\sigma(r_\theta(x,y_i) - r_\theta(x,y_{1-i})))]$$

- $r_\theta(x,y)$ is the scalar out pout of the reward model for post x, summary y with parameters θ, and D is the dataset of human judgments

Step 3: Optimize a policy against the reward model



- Reward 모델의 logit output을 강화학습을 위한 reward로 처리하는 과정. Proximal Policy Optimization(PPO) 알고리즘 사용
- 전체 보상 R(x,y)는 아래 식과 같음

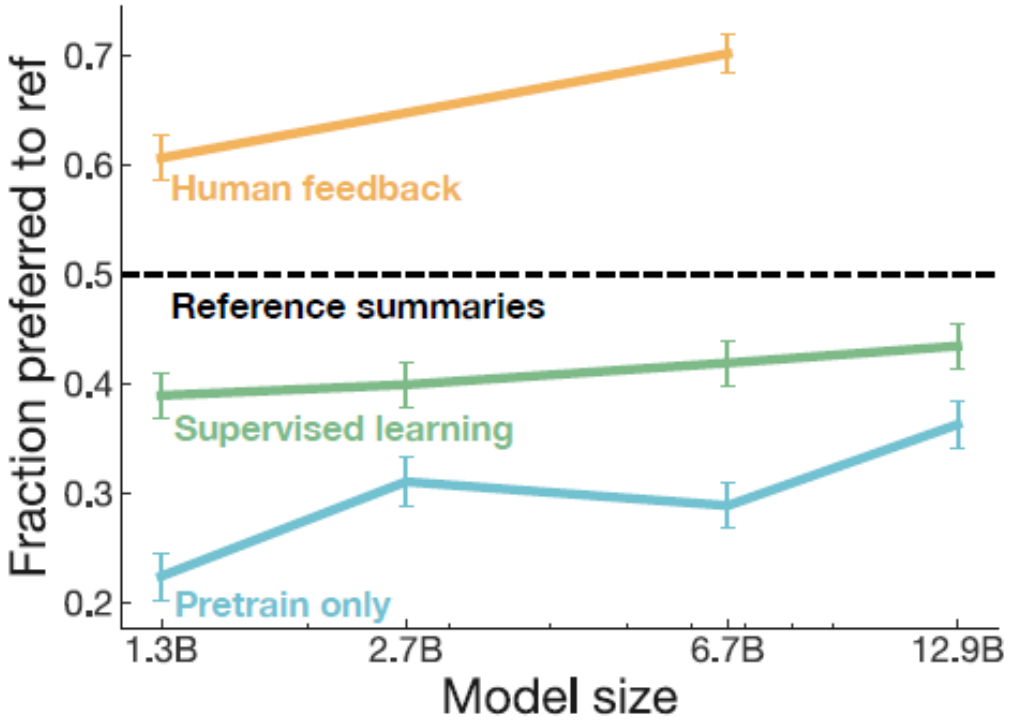
$$R(x, y) = r_\theta(x, y) - \beta \log[\pi_\phi^{\mathrm{RL}}(y|x)/\pi^{\mathrm{SFT}}(y|x)]$$

- 보상에는 학습된 정책 $\pi_\emptyset^{RL}$, TL;DR dataset에서 supervised learning으로 fine-tuning한 model π^{SFT} 간 Kullback-Leibler(KL) divergence penalty 부여

(학습 과정에서 발생할 수 있는 불안정성 줄이고, 초기에 학습된 유용한 행동 패턴 유지)

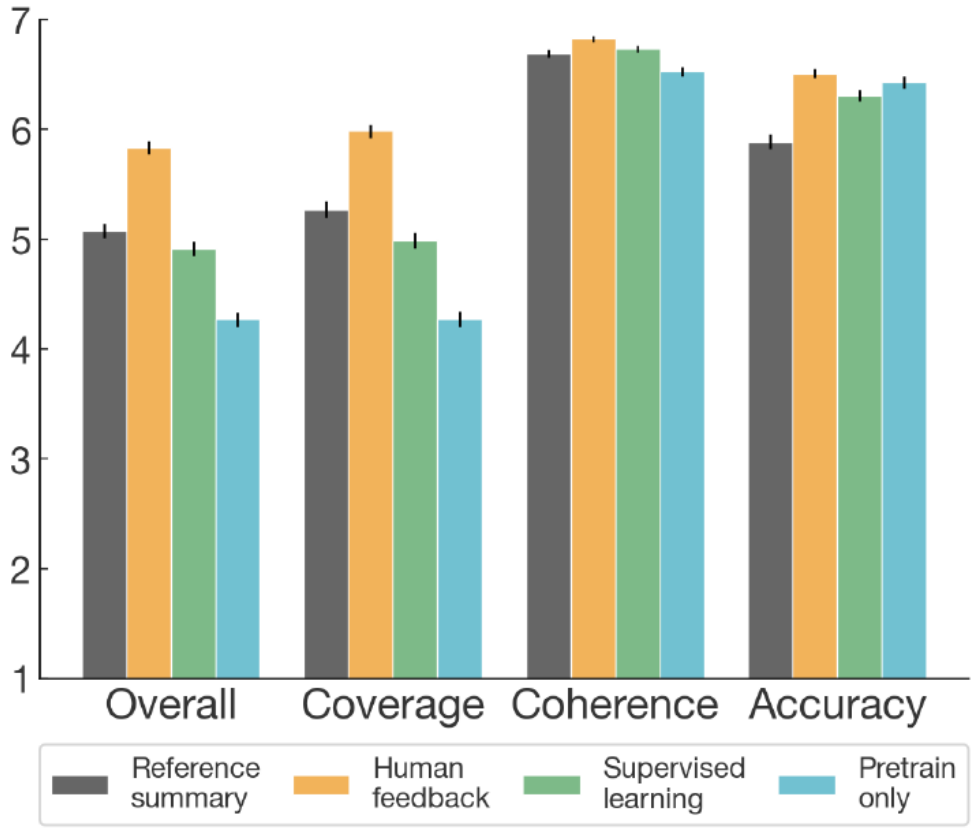
Summarizing reddit posts from human feedback

- TL;DR 데이터셋에서 사람이 생성한 참조 summary 보다 제안된 모델의 summary를 선호하는 시간의 비율
- Human feedback으로 학습된 정책들은 훨씬 큰 supervised 정책들보다 선호 됨



Summarizing reddit posts from human feedback

- 7점 리커트 척도 사용하여 4개 차원에서 summary quality 평가
- Human feedback model은 모든 차원에서 baseline models를 능가하지만, 특히 coverage에서 두드러짐

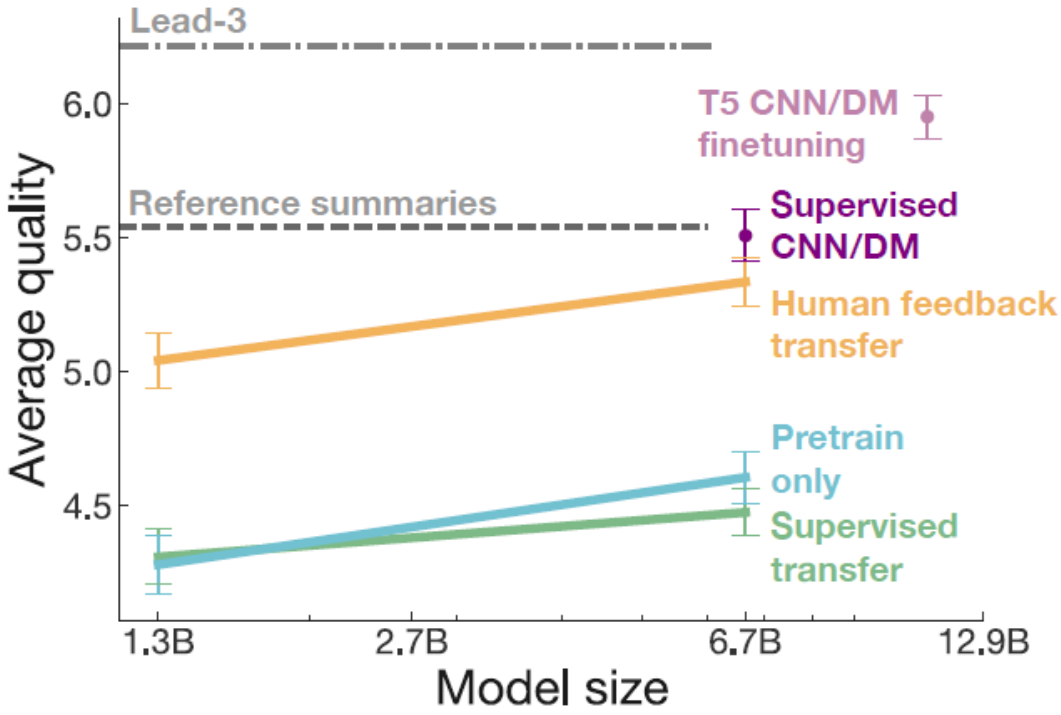


- Overall : 전반적인 summary quality
- Coverage : summary가 원본 텍스트의 주요 내용과 주제를 얼마나 포괄적으로 다루고 있는지를 나타냄
- Coherence : summary의 논리적 일관성과 구조적 흐름을 의미함
- Accuracy : summary가 원본 텍스트의 사실과 정보를 얼마나 정확하게 전달하는지를 나타냄

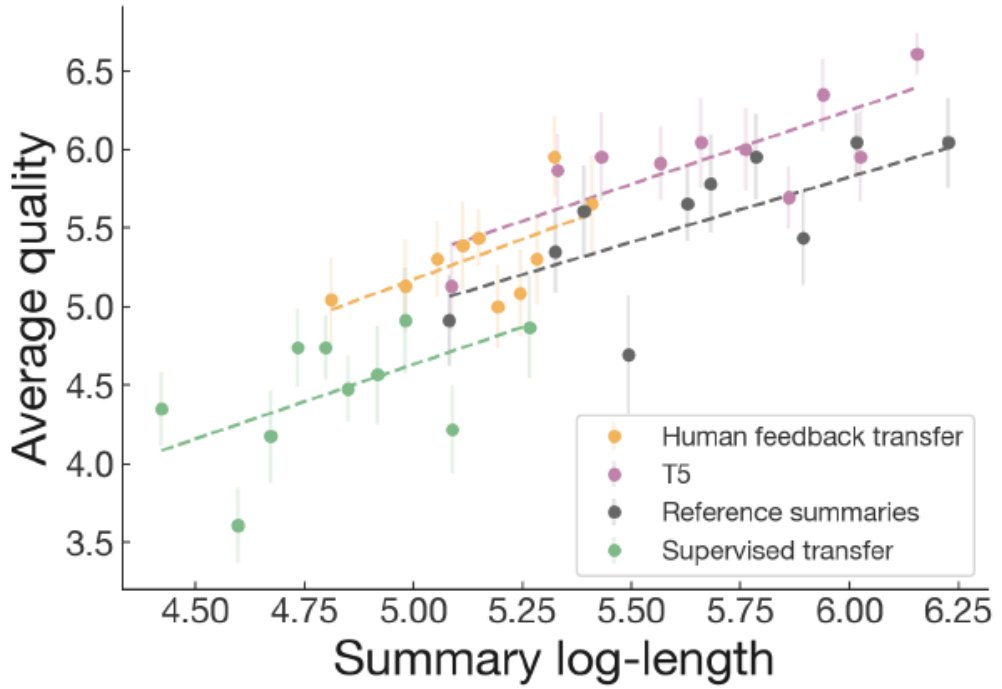
Transfer to summarizing news artides

- Transfer results on CNN/DM
- (왼) human feedback 모델은 CNN/DM 데이터에 대해 fine-tuning이 이루어진 6.7B 모델만큼 성능이 좋음
- (오) summary length 분포가 거의 겹치지 않아 직접 비교하기는 어려움

model size에 따른 summary quality



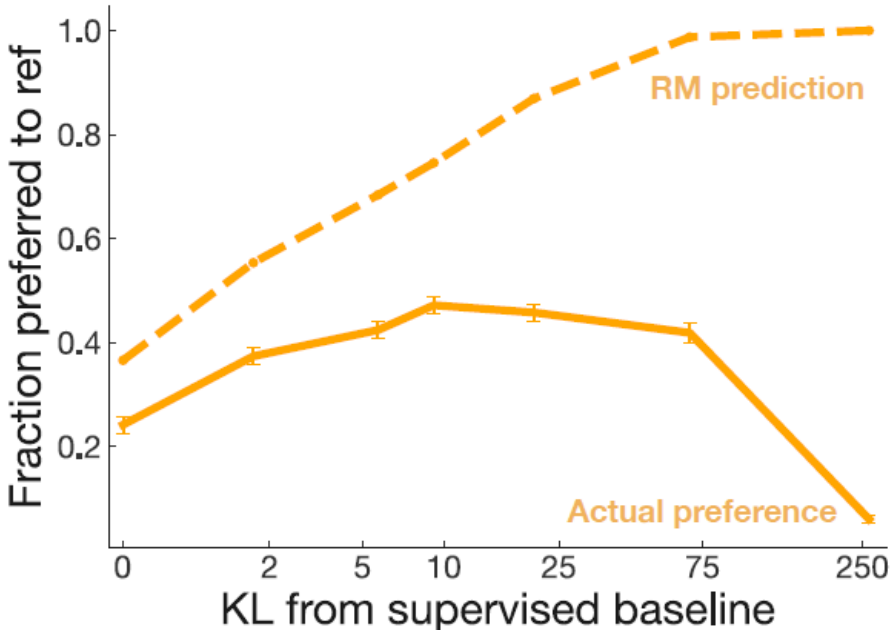
생성된 summary length에 따른 summary quality



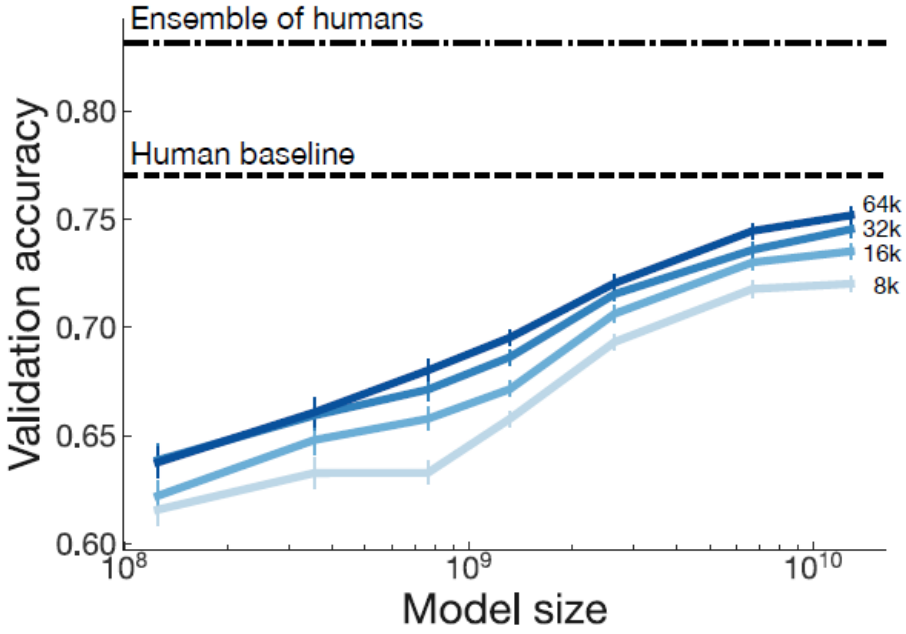
Understanding the Reward Model

- (왼) the results for PPO at a range of KL penalty coefficients
- 최적화를 거치면 모델이 개선되지만, 더 많이 최적화할수록 선호도는 예측에 비해 감소하고, reward model은 사람의 선호와 멀어짐

Reward model 최적화 정도 대비 preference scores



데이터 크기 및 모델 크기에 따른 reward model 성능



Conclusions

- 최종 모델을 생산하기까지 시간과 비용이 많이 듬
- Transfer learning에 비해 데이터 수집 비용이 많이 듬 (사람이 labeling하기 때문에 고품질의 데이터 양산 힘듬)
- 모델 출력의 quality를 평가하는 모든 task에 응용 가능함

Q & A