Recursively Summarizing Books with Human Feedback (2021)

Wu, Jeff, et al. arXiv preprint arXiv:2109.10862 (2021). Open AI

IIE8557-01 고등강화학습

경영과학연구실 이태헌

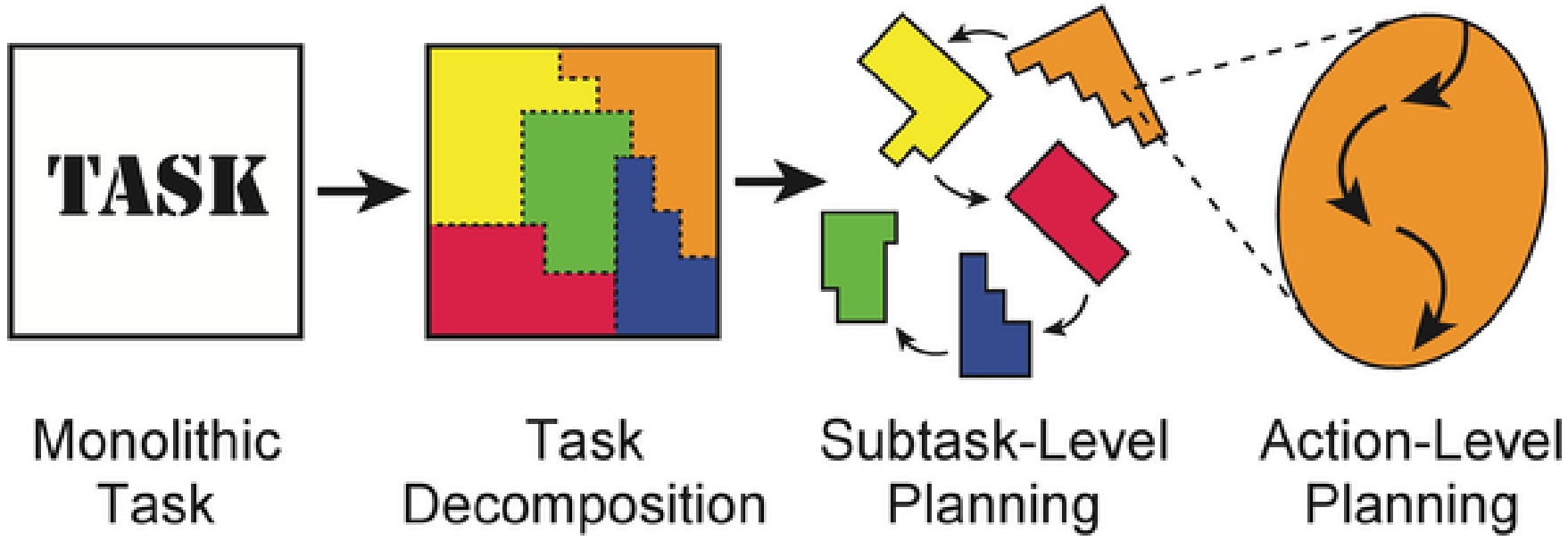
책 요약이 어려운 이유

- **책 요약은 사람이 평가하기 어렵거나 시간이 많이 소요되는 작업임. 이러한 task에 모델을 훈련 시키는 것은 한계점이 있음**

- **내용의 복잡성 : 서술적 책은** 스토리 라인, 다양한 등장인물, 배경 설명 등이 포함됨. 이 모든 것을 요약하는 것은 모델이 내용을 정확하게 파악하고 중요한 요소를 추출할 수 있어야 함
- **Human feedback 어려움**: 책 요약은 human feedback없이 모델의 정확도와 품질을 보장하기 어려움. Human feedback이 필요하지만, 사람이 짧은 시간 안에 책 전체를 읽고 요약하는 것은 많은 시간이 필요함
- **정확도와 품질의 균형**: 좋은 요약은 원본의 핵심 내용을 축약하면서도 그 의미와 맥락을 정확하게 전달해야함

Task decomposition

- 큰 문제를 더 작고 쉬운 task로 나누는 방법
- 복잡한 task를 단계별로 분해하여 각각의 작은 task를 해결하고, 결과를 통합하여 전체 문제를 해결함
- 논문에서는 책 전체를 요약하는 대신, 책을 여러 부분으로 나누어 각 부분의 요약들을 통합하여 전체 책의 요약을 생성함

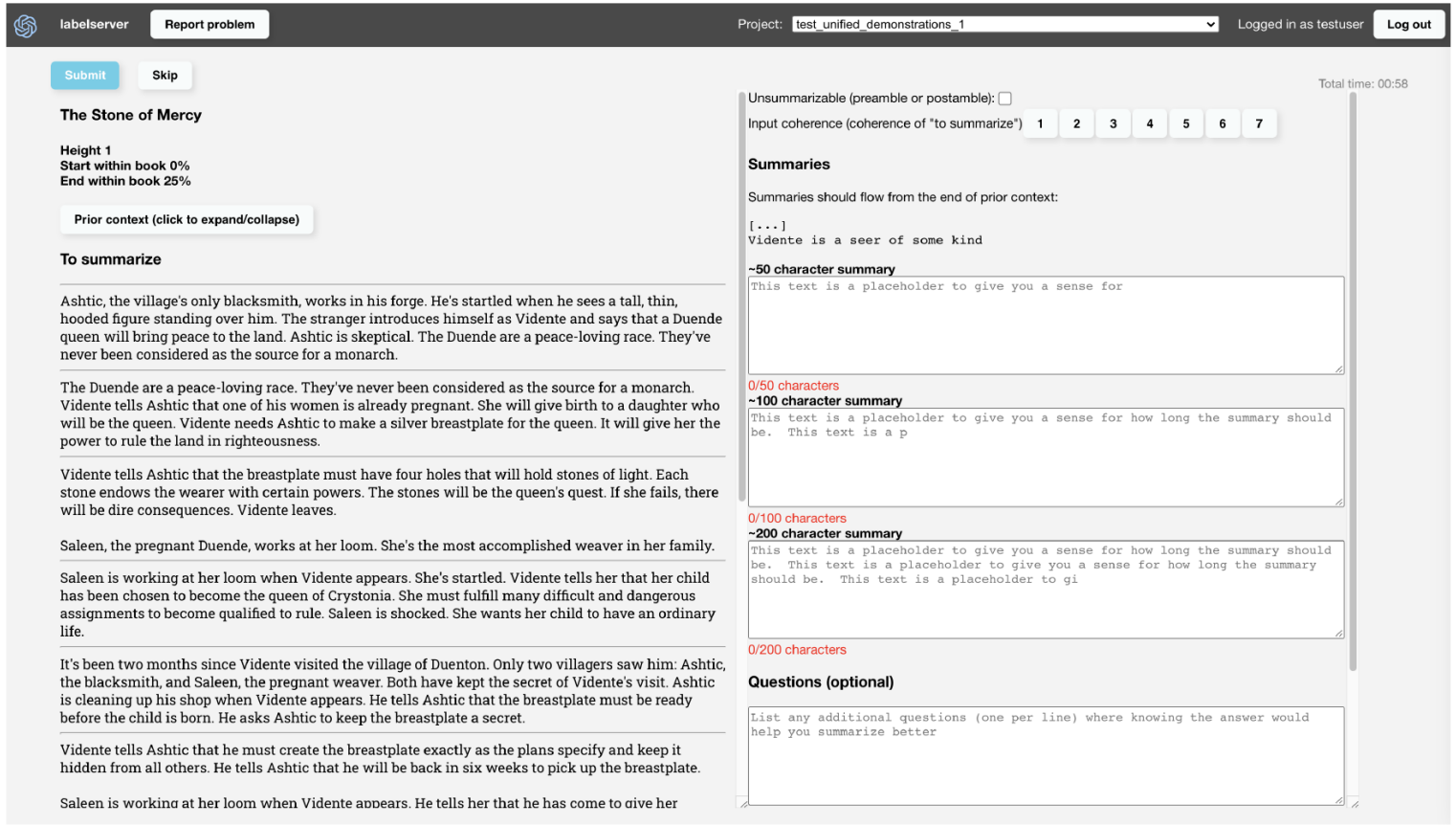


‘Human feedback을 사용한 책 요약’

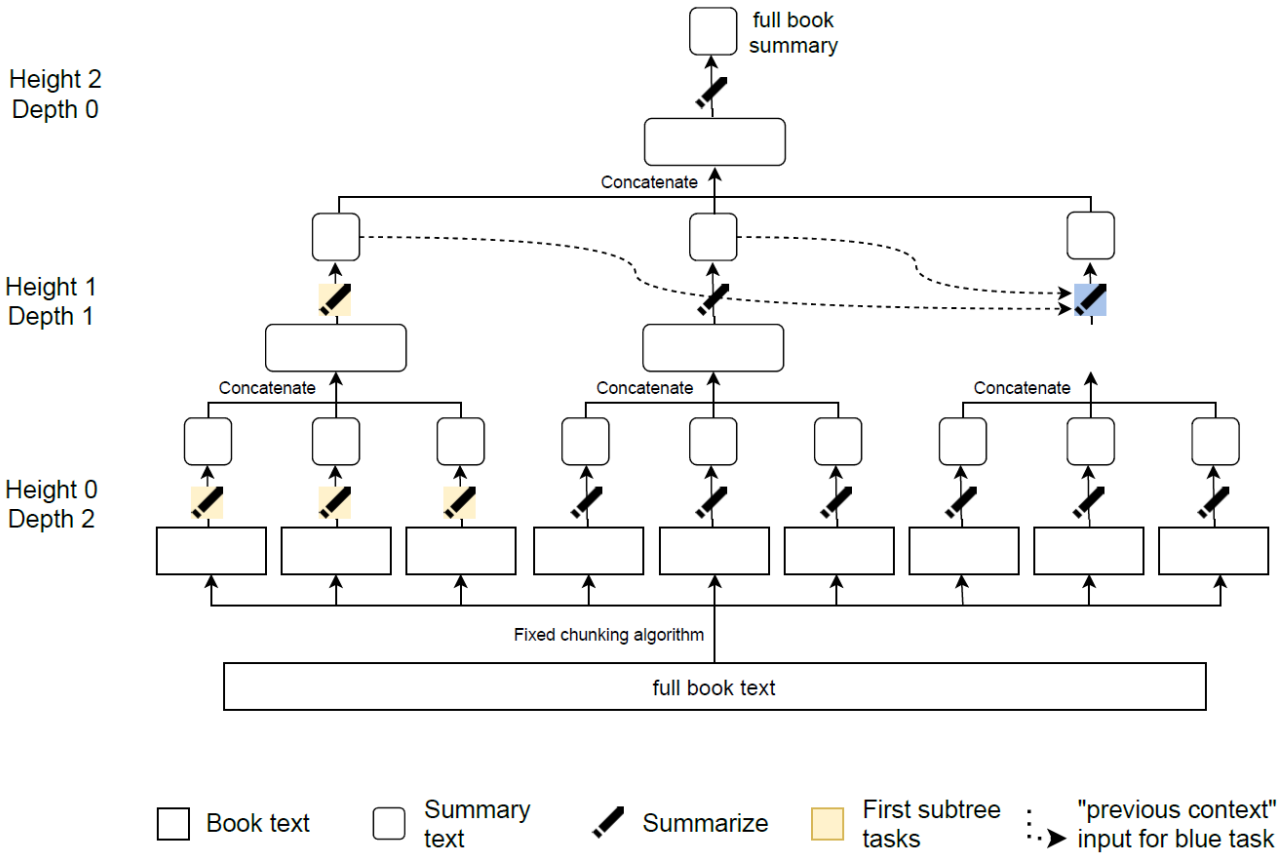
- Human Feedback 활용하여 책의 추상적 요약 수행
- Task decomposition 방법으로 한 권의 책 세분화
- 세분화 된 부분들을 Human Feedback과 Recursive 요약 과정 과정 통해 전체 책에 대해 요약
- 사람이 섹션을 직접 읽고 평가하는 대신, 모델에 의해 생성된 요약을 평가함으로써 사람의 시간을 절약하면서도 품질 높은 요약을 생성할 수 있도록 함

Dataset

- 평균적으로 100K 이상의 단어를 포함하는 서술적인 책(소설)으로 모델 학습
- 논문에서 사용한 Human feedback interface 예시



Procedure for summarizing books that combines task decomposition with learning from human feedback



1. Decomposition : 책 전체 텍스트를 여러 부분으로 나눔

2. Summarization : 각 분할된 섹션을 개별적으로 요약

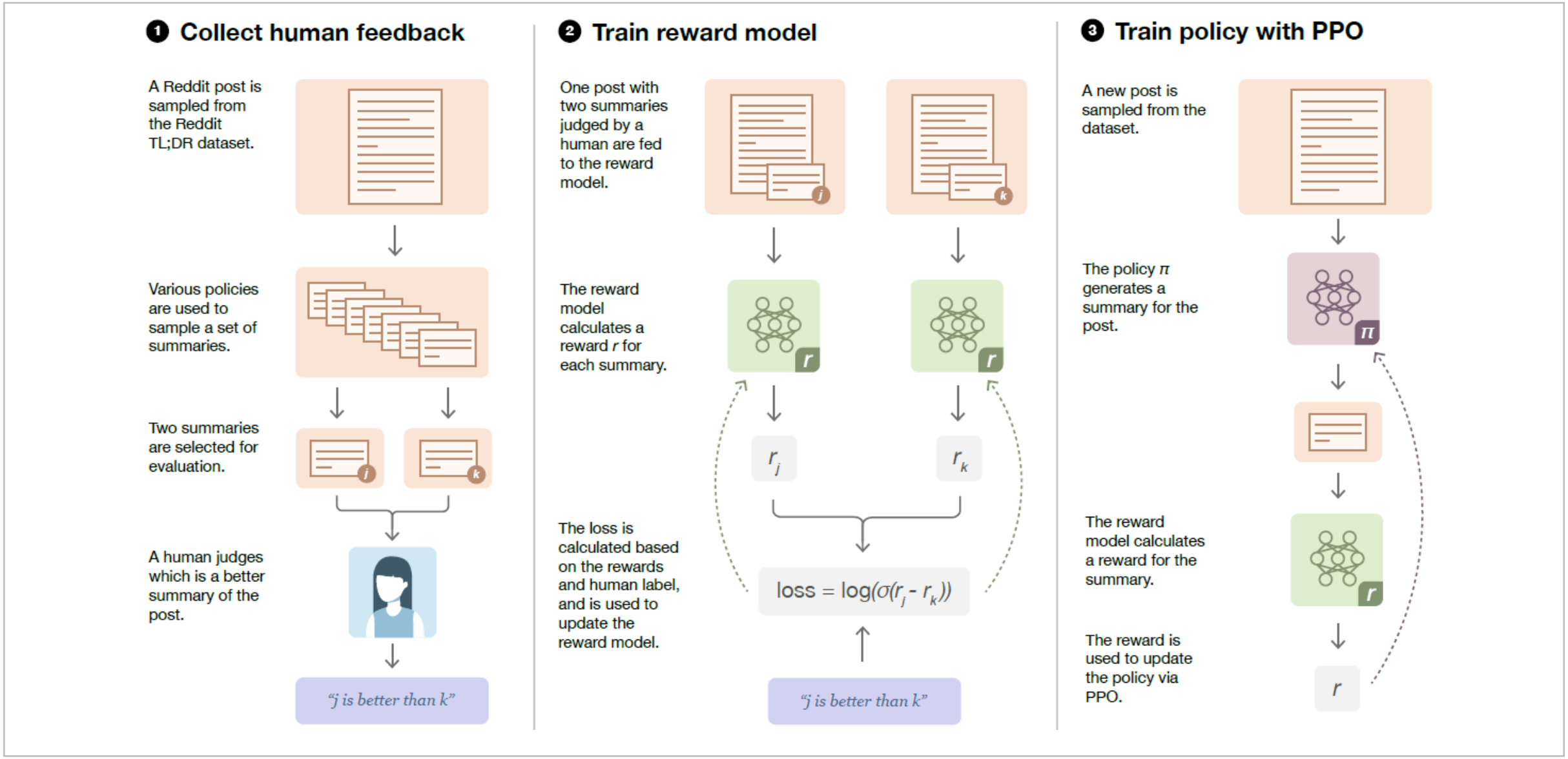
3. Feedback Collection : 평가자들이 각 요약을 평가하고, 피드백 제공

4. Behavioral Cloning : 평가자들의 피드백을 기반으로 모델을 훈련시키고, 요약 성능 향상

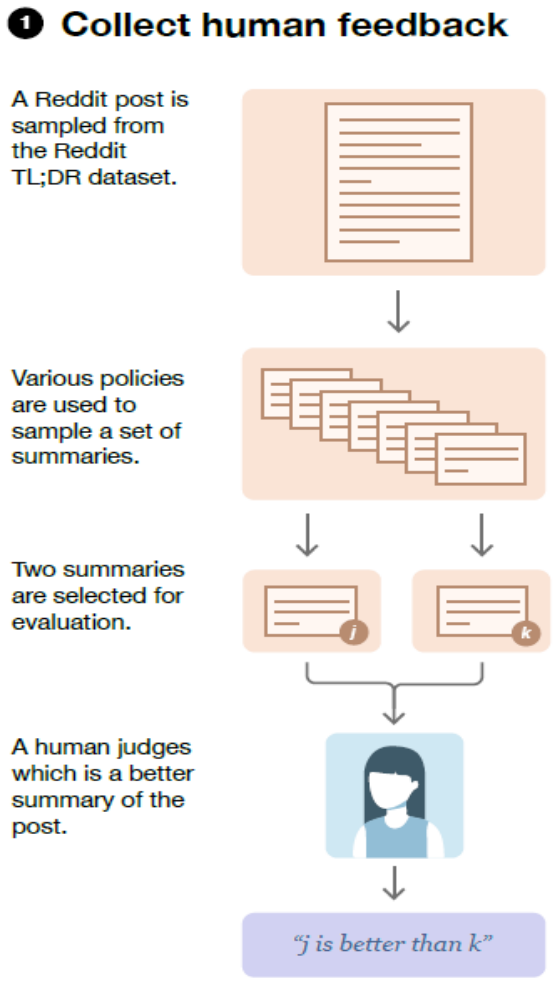
5. Reward Modeling : 모델이 생성한 다양한 요약 간의 비교를 통해, 더 좋은 요약으로 평가된 피드백 통해 모델 fine-tuning

6. Recursive Summarization : 생성된 요약들을 합쳐 더 큰 섹션의 요약을 만들고, 이 과정을 반복하여 전체 책의 요약 생성

learning to summarize with human feedback과 동일한 Training process

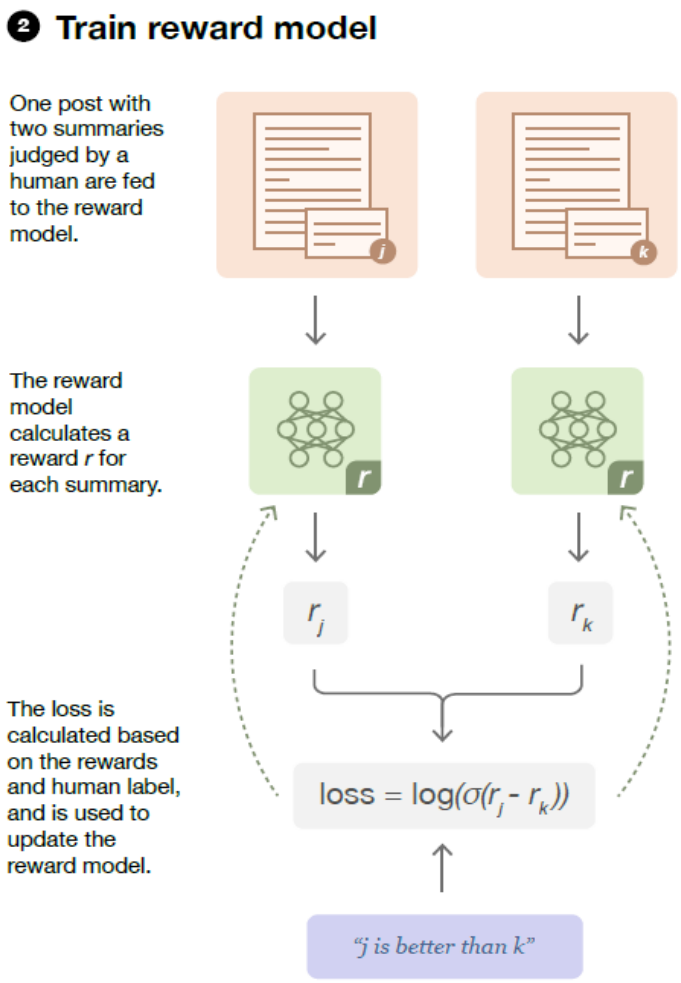


Step 1 : collect samples from existing policies and send comparisons to humans



- 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



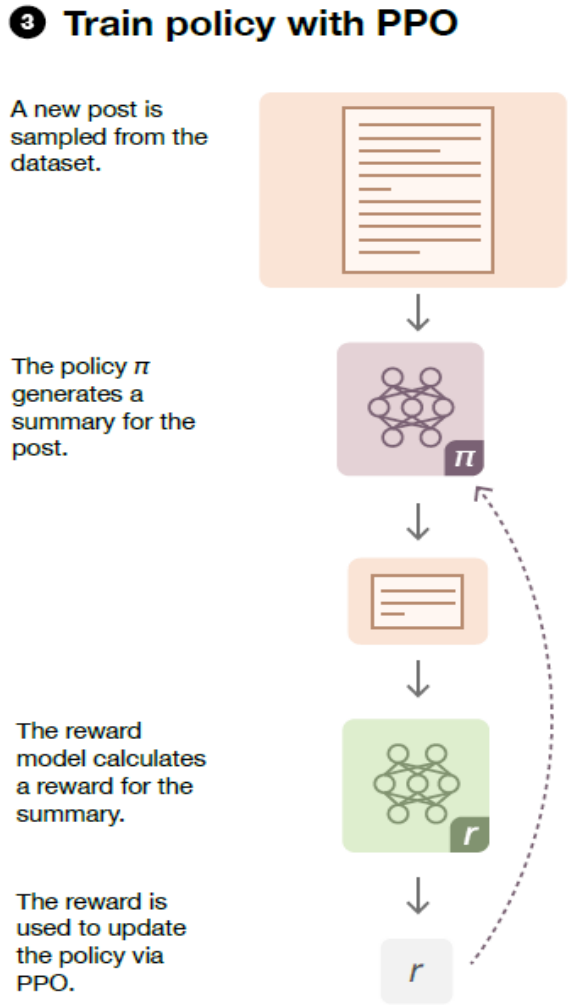
- Post와 후보 summaries가 주어지면, reward model이 summary가 human evaluator의해 더 낫다고 판단될 로그 확률을 예측하도록 학습됨
- 모델은 post x가 주어졌을 때, summary y $\in \{y_0, y_1\}$ 중 어느 것이 사람에 의해 더 낫다고 평가되는지 예측
- Human evaluator가 선호하는 summary가 y_i인 경우, reward model의 loss function은 아래 식과 같음

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

σ : $sigmoid\ function$ (0,1) 범위 출력. 두 요약의 점수 차이가 크면 1에 가까운 값, 차이가 작거나 음수이면 0에 가까운 값 출력함

- $r_\theta(x,y)$ 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

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

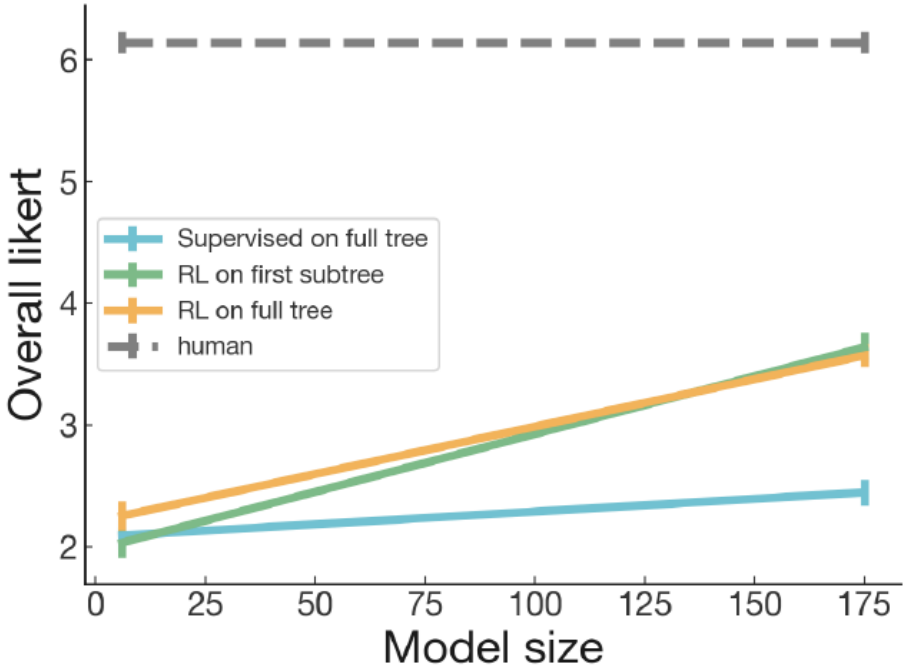
- 보상에는 학습된 정책 $\pi_\emptyset^{RL}$, TL;DR dataset에서 supervised learning으로 fine–tuning한 model π^{SFT} 간 Kullback-Leibler(KL) divergence penalty 부여

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

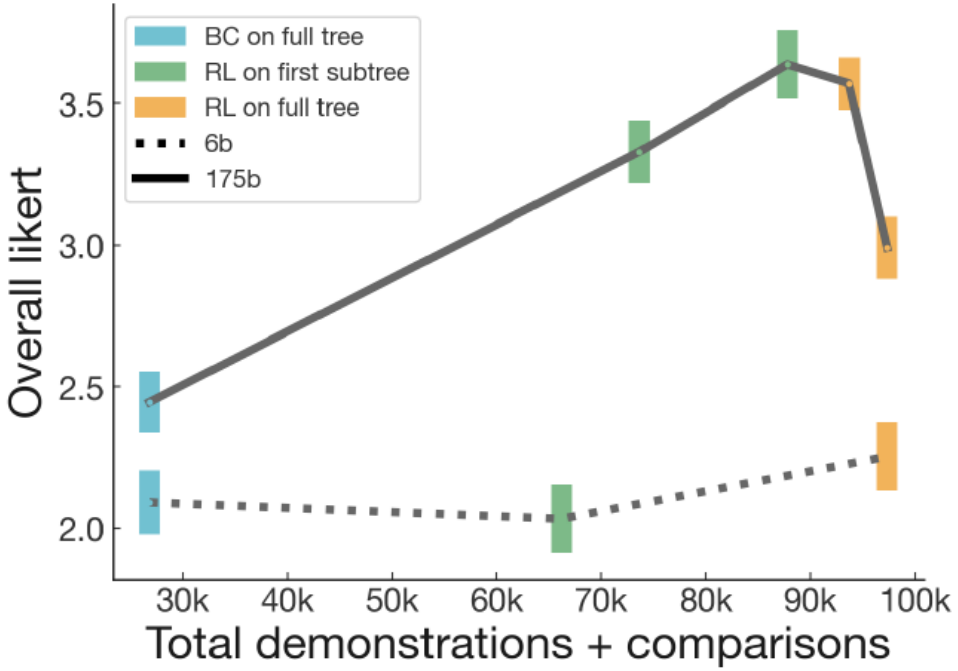
Result on full book evaluations

- 각 책에 대한 요약 평점을 1-7 리커트 척도로 평가 후 결과 비교
- RL 모델들은 BC 모델 성능을 능가하지만, 인간의 요약 성능에는 크게 미치지 못하였음

모델 크기에 따른 결과



데이터 크기에 따른 결과



- Demonstrations : 라벨러가 작성한 요약 (높은 점수의 요약본)
- Comparisons : 두 가지 모델 요약 중 어느 것이 나은지 라벨러가 비교하여 평가한 데이터

BookSum results

- 책 요약 위해 제안된 BookSum 데이터셋에서 모델 평가 결과

	Abstractive	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
Extractive Oracle		**46.62**	9.17	**18.31**	0.082
BertExt		36.71	6.16	13.40	0.028
T5 zero-shot	✓	35.43	5.62	12.02	0.011
T5 fine-tuned	✓	39.46	7.69	13.77	0.060
175b full tree RL	✓	41.51	10.46	16.88	**0.1821**
175b first subtree RL	✓	43.19	**10.63**	17.10	0.1778
6b full tree RL	✓	36.79	7.22	14.84	0.1246

- Extractive Oracle : 요약 작업에서 이상적인 추출적 요약 성능 나타내는 기준점. 최적의 추출적 요약 성능을 나타내는 이론적인 상한선
- ROUGE : 요약이 원문의 중요한 부분을 얼마나 잘 포함하고 있는지에 대한 평가 지표
  (1 : 요약과 원문 사이 단어, 2 : 이웃하는 단어 쌍, L : 요약과 원문 사이 가장 긴 연속적 일치 순서)
- BERTScore : 요약과 원본이 유사한 의미를 가지는지에 대한 평가 지표

NarrativeQA results

- NarrativeQA : 책, 영화에 대한 질문/답변 쌍으로 구성된 데이터셋
- 요약이 원문에 대한 질문에 답할 수 있는지 테스트 함
- 요약 모델이 질문 답변을 위해 훈련되지 않았음에도 좋은 결과를 보임

	ROUGE-L	BLEU-1	BLEU-4	METEOR
BiDAF (Kočiskỳ et al., 2018)	6.2	5.7	0.3	3.7
BM25 + BERT (Mou et al., 2020)	15.5	14.5	1.4	5.0
RoBERTa (Zemlyanskiy et al., 2021)	18.0	18.0	2.6	5.4
ETC (Zemlyanskiy et al., 2021)	18.8	17.2	2.7	5.4
ReadTwice (Zemlyanskiy et al., 2021)	23.3	21.1	4.0	7.0
Retriever + Reader (Izacard and Grave, 2020)	32.0	35.3	7.5	11.1
175b full tree, depth 1	21.03	21.82	3.87	10.52
6b full tree, depth 1	17.01	19.09	2.75	8.53
175b ***first subtree***, depth 1	21.55	22.27	4.24	10.58
175b full tree, ***depth 0***	18.47	20.29	3.16	9.04

- BLEU : 생성한 답변과 실제 답변이 단어단위 (1, 4)에 대해 얼마나 일치하는지 평가
- METEOR : 생성한 답변이 실제 답변과 의미적으로 얼마나 일치하는지 평가

Conclusions

- 전체 책에 대한 설득력 있는 요약을 반복적으로 생성할 수 있는 모델을 제안함
- 제안 모델을 통해 임의의 길이의 책을 요약할 수 있으며 Supervised Learning과 비교하였을 때 좋은 요약 성능을 보임
- 책 요약을 생성하거나 평가하기 위해서는 사람이 전체 책을 읽어야 하므로, 데이터셋 수집에 많은 비용이 발생함
- 장문의 NLP task에 대해 human feedback에서 학습, 작업 분해를 결합하는 것이 실용적인 접근 방식이 될 수 있음을 시사함

Q & A