



## 저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

碩士學位 請求論文

指導教授 金 壯 炫

도메인 데이터를 활용한  
응답 생성 모델 강화학습 적용

成均館大學校 情報通信大學院

빅데이터 學科

閔 丙 哲

碩  
士  
學  
位  
請  
求  
論  
文

도메인 데이터 활용 응답 생성 모델  
강화학습 적용

2  
0  
2  
4

閔  
丙  
哲

碩士學位 請求論文

指導教授 金 壯 炫

도메인 데이터를 활용한

응답 생성 모델 강화학습 적용

Application of Reinforcement Learning to Response  
Generation Model Using Domain Data

成均館大學校 情報通信大學院

빅데이터 學科

閔 丙 哲

碩士學位 請求論文

指導教授 金 壯 炫

도메인 데이터를 활용한  
응답 생성 모델 강화학습 적용

Application of Reinforcement Learning to Response  
Generation Model Using Domain Data

이 論文을 工學 碩士學位請求論文으로 提出합니다.

2023 年 10 月 日

成均館大學校 情報通信大學院

빅데이터 學科

閔 丙 哲

이 論文을 閔丙哲 의 工學  
碩士學位 論文으로 認定함.

2023 年 12 月 日

審査委員長

---

審査委員

---

審査委員

---

# 목차

제 1 장 서론.....	1
제 2 장 이론적 배경 .....	4
2.1 관련 연구 .....	4
2.2 BERT (Bidirectional Encoder Representation from Transformers) .....	7
2.3 GPT (Generative Pre-Training Transformer) .....	9
2.4 PPO (Proximal Policy Optimization) .....	10
제 3 장 연구 방법.....	12
3.1 제안 모델 .....	12
3.2 데이터 수집 및 전처리 .....	18
3.3 데이터 구성 .....	20
3.4 모델 성능 평가 .....	21
제 4 장 실험 및 결과.....	24
4.1 모델 학습 .....	24
4.2 실험 결과 .....	28
제 5 장 결론 및 향후 연구.....	30

## 표목차

[표 1] 웹 크롤링 데이터 전처리 예시 .....	19
[표 2] 학습을 위한 데이터 구성 .....	20
[표 3] 생성 모델 학습 데이터 예시 .....	20
[표 4] 모델 별 추가 학습에 사용된 하이퍼파라미터 .....	24
[표 5] 학습에 사용된 환경 정보 .....	25
[표 6] 학습 모델 검증 결과 .....	28
[표 7] 질문에 해당하는 학습 모델 별 생성 데이터 비교 예시 .....	29



## 그림목차

[그림 1] SELF-ATTENTION 의 Q, K, V 계산 .....	8
[그림 2] GPT 모델 학습 과정 .....	13
[그림 3] LOGSIGMOID 함수.....	14
[그림 4] 모델 구조 및 학습 과정 .....	15
[그림 5] 강화학습 지표 .....	27

## 논문요약

### 도메인 데이터를 활용한 응답 생성 모델 강화학습 적용

ChatGPT 공개 이후로 많은 사람들이 대규모 언어 모델의 성능을 체감하고 다양한 활용 방법을 SNS 를 통해 공유하고 있다. 대부분 In-Context Learning 을 활용하는 방법으로 특정 목적을 위해 사용하기 위해서는 추가 데이터로 학습된 모델이 필요하다. 하지만 추가 학습된 모델을 만들기 위해서는 전문 인력, 인프라, 시간, 비용 등 현실적으로 어려운 부분이 존재한다. OpenAI 에서 GPT 모델을 추가 학습하는 API 를 제공하지만 기업 내부 데이터를 외부에 유출해야 하기 때문에 보안에 위배되는 문제가 있다.

본 연구에서는 비즈니스의 특정 목적을 달성하기 위해 ChatGPT 학습 방식을 적용한 생성 모델 구축 방안을 제시하고 성능을 확인한다. 의료 분야 상담 데이터를 수집하여 환자들의 질의에 응답을 생성하는 과제로 연구의 범위를 한정했다. 또한 추가 학습 모델이 생성한 여러 응답 중 사람이 직접 선호하는 문장에 순위를 정하는 HumanFeedback 방식 대신 전문의 활동 정도를 나타내는 점수를 기반으로 순위를 정하였다.

추가 학습을 진행한 모델과 강화학습을 적용한 모델이 생성한 텍스트를 각각 전문의 응답과 유사도를 측정하였고 강화학습을 적용한 모델이 성능이 향상된 것을 확인할 수 있었다. 강화학습 과정에서는 보상의 편차가 심하였지만 여러 지표에서

학습이 올바르게 진행됨을 확인하였다. 또한 응답을 확인 했을 때 추가 학습 모델은 반복적인 단어를 생성하거나 문장의 끝맺음이 되지 않은 경우가 많은 반면에 강화학습된 모델에서는 이러한 부분이 개선됨을 확인 하였다.

주제어 : Generative Model, LLM, Reinforcement Learning, Q&A, Natural Language Preprocessing

# 제 1 장 서 론

딥러닝의 발전으로 다차원의 데이터 처리 방식이 대중화 되었으며, 이를 통해 사람의 다양한 표현을 컴퓨터가 이해하도록 기술적으로 구현 가능하게 되었다. 대화시스템 분야에서는 문맥, 문장에서 단어의 의미, 문장의 의도 등 대화에 필요한 정보를 다차원 데이터에 기록하고 이를 적절히 처리하며 사람과의 유연한 소통을 가능하게 했다. 대화 시스템은 응답하는 주제에 따라 오픈 도메인 대화 시스템 (Open Domain Dialogue System)과 클로즈 도메인 대화 시스템 (Close Domain Dialogue System)으로 나눈다. 오픈 도메인 대화 시스템은 자유 주제에 대하여 사용자와 의사 소통을 하는데 목적이 있으며, 클로즈 도메인 대화 시스템은 특정 주제에 대해 자세한 응답을 제공 하거나 예약/신청/취소 등과 같이 사용자 의도를 달성하는 것을 목적으로 한다.

최근 트랜스포머(Transformer) [1]를 기반으로 한 대규모 모델 LM(Large Model)이 출현하며 많은 관심을 받고 있다. 트랜스포머의 병렬 구조는 큰 규모의 모델 학습을 가능하게 했고, 자연어 처리 분야에서는 LLM (Large Language Model)과 같이 많은 양의 텍스트 데이터를 학습한 모델이 오픈 소스로 공개되었다. 대표적인 LLM 으로 GPT3(Generative Pre-trained Transformer 3) 모델[2]이 있으며 막대한 양의 데이터와 175B 규모의 파라미터를 가졌다. GPT3 는 다음 토큰을 예측하는 형태로 학습 되었으며 파라미터 규모가 커질수록 놀라운 텍스트 생성 성능을 보였다. 이러한 성과로 대화 시스템에도 사용자 질문이 입력되면 답변을 바로 생성하는 생성 기반 대화 시스템 연구가 진행되고 있다.

대화 시스템의 전통적인 구조는 발화를 이해하는 NLU(Natural Language Understanding), 슬롯의 값을 추적하는 DST(Dialogue State Tracking), 시스템 액션을 결정하는 DPL(Dialogue Policy Learning), 시스템 응답을 생성하는 NLG(Natural Language Generation) 와 같은 4 개 모듈의 조합으로 이루어졌다. 하지만 이러한 방식은 각 모듈을 독립적으로 최적화 시켜야 하는 어려움이 있다. 또한 각 모듈을 학습시키는 데이터를 구축하기 위해 데이터 라벨링에 많은 비용이 발생한다. 생성 모델 기반 End-to-end 방식의 연구[3, 4] 는 기존 4 개 모듈을 하나의 모듈로 줄여 기존의 어려움을 해결하였다. 그리고 [usr], [sys], [ds] 등과 같은 스페셜 토큰을 사용하여 모델을 추가 학습 시키는 방법은 사전 학습된 모델과의 불일치를 유발할 수 있다. 그래서 User, System 과 같이 스페셜 토큰을 추가하여 학습하기 보다는 프롬프트에 텍스트 형태로 모델에 입력하는 연구[5]도 제안되었다.

여러 생성 기반 모델 대화 시스템의 연구에도 불구하고 실제 고객 서비스를 위해 기업에서 적용하기에는 리스크가 존재한다. 생성 모델은 Cross Entropy 손실 함수를 사용해 다음 토큰을 예측 하도록 훈련 되었기 때문에 편향되거나 사용자 의도와 다른 정보를 생성(Hallucination)하는 단점이 있다. 사용자 의도에 맞지 않는 정보를 제공하는 것은 비즈니스에 큰 악영향을 끼친다. 이러한 Hallucination 을 보완하기 위한 대표적인 방법으로는 유사도 기반으로 검색한 텍스트로 학습 시킨 생성 모델 RAG(Retrieval-Augmented Generation) [6]와 사람이 평가한 보상으로 강화 학습한 RLHF(Reinforcement Learning from human feedback) 가 있다.

인간이 선호하는 답변을 얻기 위해 인간이 직접 평가해야 한다는 정의 아래에 제안된 RLHF 방법으로 학습시킨 InstructGPT[7] 모델을 OpenAI 에서 발표했다. 그리고 상용화 버전으로 ChatGPT 가 출현하여 많은 사람들이 생성 기반 서비스에 대해 편리성을 인식하게 되었다. 기업에서는 ChatGPT 의 기능을 활용하여 다양한 서비스를 출시하고 있다. 하지만 사내 데이터가 ChatGPT 서버에 전송되는 보안적인 문제 때문에 활용에 있어 한계점이 존재한다. 따라서 회사에서 보유한 데이터로 생성 모델을 학습하고 정제하는 방법론이 필요하다.

본 논문에서는 InstructGPT 학습 방법을 토대로 특정 도메인의 사내 데이터로 응답 생성 모델을 구축하는 방법론을 제안한다. 특정 분야의 도메인 데이터를 수집하기 위해 웹크롤링을 하였고 헬스케어 상담 Q&A 데이터를 수집했다. 먼저 SFT(Supervised fine tuning) 단계에서 수집한 데이터로 기본적인 생성 모델을 학습하였다. 이후 모델 출력을 정제하기 위해 출력한 응답의 보상 값으로 파라미터를 조정하는 강화학습을 적용하였다. 최종 모델의 성능 측정은 문맥적 점수를 나타내는 BERTScore 지표를 활용하였고, 강화학습된 모델이 SFT 모델에 비해 질문의 목적에 맞는 응답과 사람이 선호하는 답변을 모델이 출력하는지 확인했다. 본 연구를 통해 기여한 바를 다음과 같이 요약할 수 있다.

- 한국어 헬스케어(의료 상담) 데이터로 LLM 학습 및 강화학습 적용
- 비용 효율적이며 외부 도구를 사용하지 않고 보상 모델을 구현

## 제 2 장 이론적 배경

제 2 장에서는 환자와 전문의 상담 Q&A 데이터를 적용한 자연어 생성 모델을 InstructGPT 방식으로 학습하는 연구를 제안한다. 이 장에서는 자연어 생성 연구 동향과 연구에 사용된 모델에 대해 살펴본다.

### 2.1 관련 연구

자연어 생성 분야에서는 딥러닝 기술을 이용한 다양한 언어 모델을 사용한다. 대표적으로 인코더-디코더 구조에 RNN(Recurrent Neural Network) [8], LSTM(Long Short Term Memory) 과 같은 연속적인 데이터를 순차적으로 처리하는 알고리즘을 사용하는 방법이 있다. 하지만 인코더는 입력 문장 전체를 고정된 길이의 벡터(Context Vector)로 축약하기 때문에 디코더를 통해 자연어를 생성할 때 정보 손실이 발생할 수 있다. 이후 인코더-디코더 구조에 Attention 알고리즘을 적용[9]하여 모든 입력 값의 히든스테이트를 사용하도록 했다. 디코더에서 토큰을 생성할 때 히든스테이트를 참고하도록 하여 문장의 길이가 길어지더라도 성능을 유지할 수 있도록 했다. 하지만 RNN, LSTM 알고리즘과 같이 연속적 데이터를 타임 스텝에 따라 순차적으로 처리하는 방식은 속도가 느리고 성능에 한계가 존재한다.

Transformer 구조는 순환신경망(RNN) 계열의 연속 데이터의 순차적인 처리 대신에 병렬 처리를 가능하게 했다. 병렬 처리로 인하여 빠른 학습이 가능해져 대규모 모델들이 출현할 수 있게 됐다. 대규모 모델 중 자연어 생성 분야에서는

Transformer 의 디코더 구조를 기반으로 GPT1[10]과 같은 사전 학습 언어 모델 PLM(Pre-trained language models)이 출시 되었다. PLM 은 전이 학습(Transfer Learning)이 가능하여 추가 학습 데이터로 미세 조정(Fine-tuning) 을 진행하면 다운 스트림 테스트에 맞는 생성 모델을 만들 수 있게 되었다. 이후 PLM 의 학습 파라미터를 증가시킨 대규모 언어 모델 LLM 이 나오며 추가 학습과 미세조정 없이도 범용적인 자연어 처리 테스트를 해결하게 되었다. LLM 성능을 확인하고 많은 기업에서 도메인 특화된 모델을 만드는 연구가 활발히 진행 중이다. Meta 에서 발표한 LLaMA (Large Language Model Meta AI) 모델[11]은 오픈소스로 제공되어 LLM 모델을 활용한 연구 개발 속도에 크게 기여하였다.

헬스케어 분야에서도 LLaMA 모델에 온라인 의료 상담 데이터를 적용하여 사용자들의 질문에 의학적인 답변을 생성하는 오픈소스 모델 연구 [12, 13, 14, 15]가 진행되었다. ChatDoctor[12] 모델의 경우 의학 상담 온라인 데이터를 수집하여 LLaMA 모델을 파인튜닝한 후 외부 지식을 적용하여 답변을 생성하였다. Med-Alpaca[13]와 PMC-LLaMA[14] 는 많은 양의 의학 서적 및 Medical NLP 테스트 데이터로 파인 튜닝하여 도메인 특화된 모델을 만들었다. 이러한 모델들은 모두 영어로 구축된 데이터를 사용하였기 때문에 다른 나라 언어로 의학 답변을 생성하는데 한계가 있다. 따라서 [15]의 연구는 중국어 데이터를 사용하여 모델을 만들어 ChatGPT 보다 성능이 높다는 것을 증명했다. 국내의 경우 단어 사전 분포만 고려하지 않고 입력 문장의 단어를 사용하여 출력하는 PGN(Pointer Generator Networks) 모델[16]과 외부 지식을 사용하여 응답을 생성하는 모델[17]을 연구했다. 이 외에 LLM 과 직접적으로 관련된 선행 연구를 찾을 수



없었다. 그 이유로는 많은 양의 한국어 의료 데이터가 공개 되어 있지 않기 때문이라고 판단된다.

본 연구에서는 GPT3 규모의 트랜스포머 기반 자연어 생성 모델을 의료 상담 데이터로 추가 학습한다. 추가 학습 만으로 특정 도메인 질문에 자연스럽게 못한 문장을 생성할 수 있기 때문에 강화학습을 적용하여 성능 향상을 확인하는 것에 초점을 맞춘다.

## 2.2 BERT (Bidirectional Encoder Representation from Transformers)

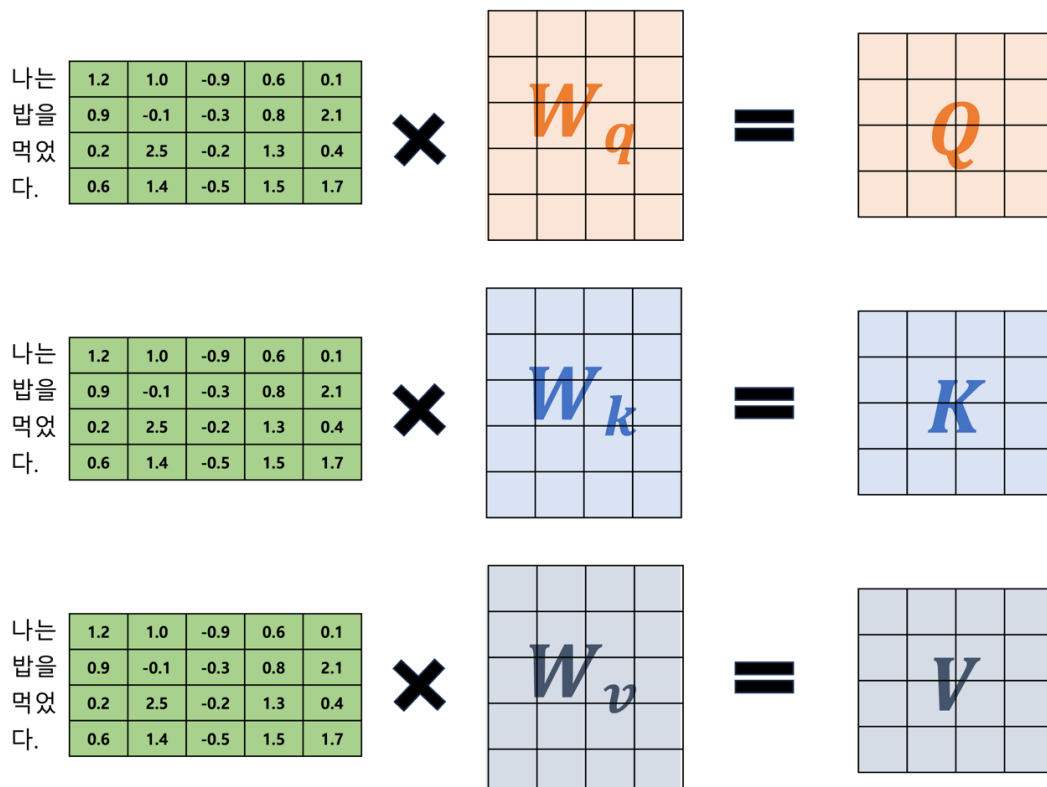
2019 년 구글에서 발표한 BERT[18] 모델은 Transformer 의 인코더 구조를 사용해서 학습한 pre-trained 모델이다. 모델 학습 시 입력 데이터는 텍스트를 숫자로 구성된 벡터로 표현해주는 Embedding 과정을 거친다. Embedding 과정은 Token Embedding 과 Segment Embedding, Positional Embedding 으로 나누어진다. 그리고 단어 사전 구축 단계에서는 BPE(Byte Pair Encoding)[19] 방식을 사용하여 단어보다 작은 단위인 서브워드(subword)를 사용했다. 서브워드 방식의 장점은 단어보다 작은 단위이기 때문에 단어 사전에 없는 단어가 출현한 경우인 OOV(Out-Of-Vocabulary) 문제를 막을 수 있다는 것이다. Segment Embedding 은 모델의 입력으로 여러 문장이 입력되는 경우 토큰이 어느 문장에 속해 있는지 정보를 주기 위한 목적으로 사용된다. 입력 문장이 두개인 경우 첫번째 문장은 0, 두번째 문장은 1 의 숫자로 표현된다. Positional Embedding 은 Sine, Cosine 주기 함수를 이용하여 토큰의 위치를 표현한다. RNN 계열의 모델은 순차적으로 데이터가 모델에 입력되어 순서 정보가 포함되지만 Transformer 계열 모델은 데이터가 병렬로 입력이 되어 순서 정보가 없다. 따라서 문장과 같이 연속성이 있는 데이터에는 순서 정보가 필요하여 Positional Embedding 을 사용한다. 최종적으로 Embedding 단계에서는 위의 세가지 Embedding 의 결과를 합하여 학습할 모델에 입력한다.

BERT 모델은 MLM(Masked Language Model)과 NSP(Next Sentence Prediction) 방식으로 사전 학습을 진행 했다. MLM 방식은 입력 데이터의 15%

토큰을 랜덤하게 선택하여 그 중 80%는 [MASK] 토큰으로 치환하고, 10%는 랜덤한 단어, 나머지 10%는 원래 단어를 사용한다. NSP 방식은 BERT 모델에 두 개의 문장을 입력하여 앞뒤 문장 간의 관계를 학습하는 것이다. 각 문장은 [SEP] 토큰으로 구분되며 연결된 문장이면 IsNext 그렇지 않으면 NotNext로 라벨링하여 학습 시킨다.

BERT 모델의 구조는 트랜스포머의 인코더를 사용한다. 인코더 내부 구조 중에 Multi-Head Self-Attention은 여러 시각에서 문장 내 단어 간 유사도를 기반으로 어느 단어에 집중해야 하는지 판단하는 알고리즘이다.

[그림 1] Self-Attention의 Q, K, V 계산



[그림 1]과 같이 Self-Attention 은 입력 데이터의 Embedding Matrix 를 학습을 통해 업데이트가 가능한 가중치인  $W_Q, W_K, W_V$  에 각각 내적 하여  $Q(Query)$ ,  $K(Key)$ ,  $V(Value)$  를 생성한다. 이후 (1)의 공식 과 같이  $Q$  와  $K^T$  를 곱하여 유사도를 계산하고 스케일링 ( $\sqrt{d_k}$ ) 후  $V$ 를 곱해 Attention 스코어를 구하여 한 개의 Self-Attention 결과를 생성한다. 최종적으로 Self-Attention 을 다수 수행하여 이어 붙인 결과와 동일한 차원의 가중치( $W$ ) 와 곱하여 Transformer 인코더의 결과를 얻는다.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

## 2.3 GPT (Generative Pre-Trained Transformer)

GPT 모델은 Transformer 의 디코더 구조를 사용해서 학습한 pre-trained 모델이다. BERT 와 동일한 임베딩 방식을 사용하며 Masked Multi-Head Self-Attention 을 사용하는 것이 BERT 와 다른 점이다. Masked Multi-Head Self-Attention 방식은 입력 시퀀스에서 입력하는 Query 다음에 오는 토큰을 모두 마스킹한다는 점에서 일반적인 Multi-Head Self-Attention 과 차이가 있다.

학습 방식은 이전 토큰의 출력을 입력하여 다음 단어를 예측하는 방식으로 진행된다. Bert 와는 다르게 현재 토큰인  $T_i$  이후 미래 토큰  $T_{i+1} \dots$  대한 연산이 필요하지 않기 때문에 마스킹 기법을 통해  $T_{i+1} \dots$  를 0 으로 치환하여 훈련 중

연산에 방해가 되지 않도록 한다. 0 으로 표현하기 위해 Attention Score 행렬에서 대각선 위의 부분을 모두  $-\inf$  값으로 변경 후 softmax 함수를 통해 Attention Score 를 모두 0 이 되도록 유도한다.

## 2.4 PPO (Proximal Policy Optimization)

PPO[20] 는 강화학습(Reinforcement Learning)의 정책 기반 알고리즘 중 하나이다. 강화학습의 목적은 에이전트(Agent)가 주어진 환경(Environment)과 상태(State)에서 기대되는 보상 값을 최대화 하는 액션(Action)의 정책(Policy)을 찾는 것이다. 강화학습 방법 중 정책 기반(Policy-based)은 수식 (1)과 같이 상태( $S_t$ )를 입력 받아 보상이 높은 액션의 확률 분포  $\pi(a|s, \theta)$ 를 출력하도록 정책 신경망 파라미터 ( $\theta$ ) 를 업데이트하는 것을 나타낸다. 파라미터 업데이트는 수식 (2)와 같이 목표 함수를 미분한  $\nabla J(\theta)$  에 학습률  $\alpha$  을 곱해 Gradient Ascent 방식으로 진행된다. 하지만 정책 신경망을 학습하는 것은 액션을 샘플링하여 가중치를 업데이트 하는 과정에서 분산(Variance)이 쉽게 커지는 단점이 있다.

$$\pi(a|s, \theta) = \mathbb{P}[A_t = a | S_t = s, \theta_t = \theta] \quad (1)$$

$$\theta_{t+1} = \theta_t + \alpha \nabla J(\theta) \approx \theta_t + \alpha \mathbb{E}_{\pi_\theta} \left[ \sum_{t=0}^T \nabla_\theta \log \pi_\theta (\alpha_t | s_t) G_t \right] \quad (2)$$

PPO 알고리즘은 분산을 줄이기 위해 클리핑(Clipping) 기법을 적용하여 기존 정책에서 새로운 정책으로 업데이트 하는 정도를 제한하였다. 과도한 업데이트를 제한하는 PPO의 목적함수는 수식 (3) 과 같다.  $r_t(\theta)$  는  $\pi\theta(a_t|s_t) / \pi\theta_{old}(a_t|s_t)$ 로 표현할 수 있으며 이전 정책  $\pi\theta_{old}(a_t|s_t)$  에 비하여 최신 정책  $\pi\theta(a_t|s_t)$ 의 확률 분포의 비율을 나타낸다. Advantage  $\hat{A}_t$  는 TD Error(Time difference Error)를 통해 구할 수 있으며 Advantage 가 높은 액션을 선택하도록 학습된다. 그리고  $\epsilon$ 은 클리핑 구간을 정하는 하이퍼 파라미터이며  $r_t(\theta)$  가  $[1 - \epsilon, 1 + \epsilon]$  구간에 비해 작으면  $1 - \epsilon$ , 크면  $1 + \epsilon$  로 클리핑 한다. 이후  $\hat{A}_t$  를 곱하여 업데이트 정도를 제한하여 안정적인 학습을 유도한다.

$$L^{CLIP}(\theta) = \mathbb{E}_t[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)\hat{A}_t)] \quad (3)$$

InstructGPT 는 GPT3 타입의 생성 모델을 추가 학습하고 강화학습 알고리즘 중 PPO 알고리즘을 적용하였다. 본 논문의 연구에서는 PPO 알고리즘을 사용하여 텍스트 생성 모델에 강화학습을 적용했으며 자세한 동작 방식은 다음 장에서 다룬다.

## 제 3 장 연구 방법

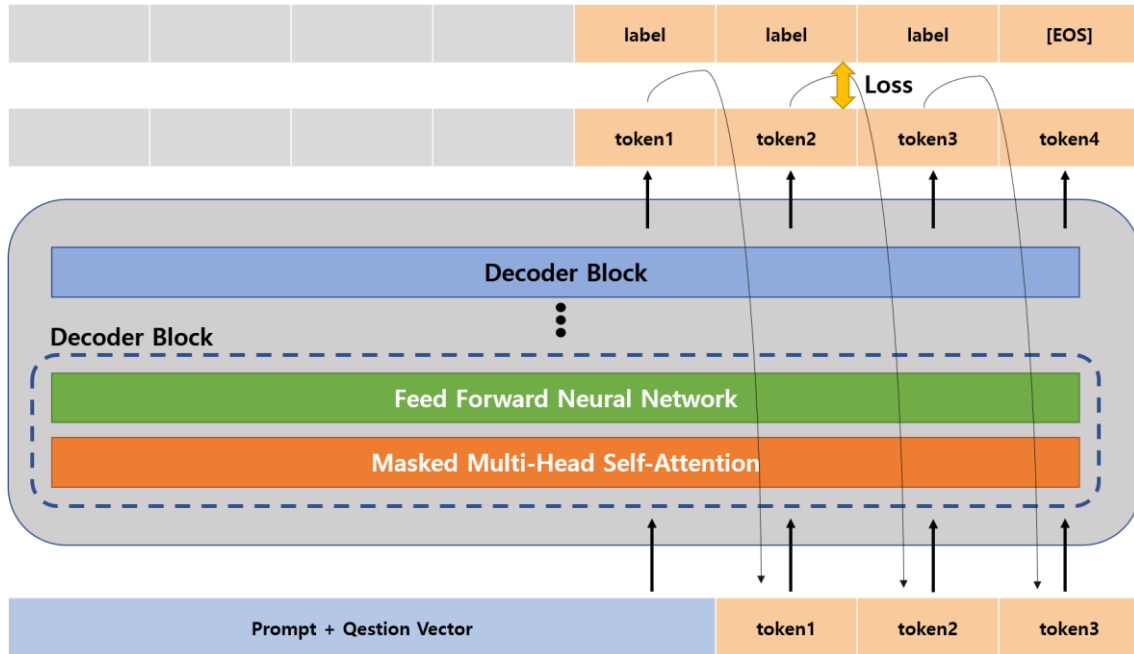
이 장에서는 직접 수집한 의료 상담 데이터로 InstructGPT 논문에서 제시한 방법으로 학습하고 응답 성능을 확인하기 하기 위한 연구를 소개한다.

### 3.1 제안 모델

전체 모델 구조는 RLHF 방법과 동일하게 SFT(Supervised Fine Tuning), RM(Reward Model), PPO(Proximal Policy Optimization algorithms) 총 세 단계로 구성된다.

SFT 단계에서는 GPT3 규모의 비영리 AI 연구단체인 EleutherAI 에서 개발한 Polyglot-ko-1.3b 모델[21]을 사용했다. Polyglot-ko-1.3b 모델은 13 억 개의 파라미터, 30003 개의 단어 사전과 한국어로 구성된 868G 훈련 데이터셋으로 사전 학습하여 만들었다. 입력 데이터는 문장의 시작을 의미하는 [BOS] 토큰 뒤에 프롬프트, 사용자 질문, 전문의 답변과 [EOS] 토큰 순서로 구성하였다. 안정적인 학습과 성능을 위해 프롬프트와 질문 부분 외에 응답 부분만 학습을 진행한다. 초기 모델에 프롬프트와 질문 벡터를 입력하면 다음 차례에 오는 토큰 확률 분포가 출력된다. 이후 확률 분포와 실제 라벨과의 손실(Loss)을 비교하여 모델의 가중치를 업데이트한다. 학습이 완료된 모델은 질문을 입력하면 모델이 [EOS] 토큰이 나올 때까지 확률을 기반한 다음 토큰을 생성한다. 모델을 학습시키는 전체적인 구조는 [그림 2]와 같다.

[그림 2] GPT 모델 학습 과정



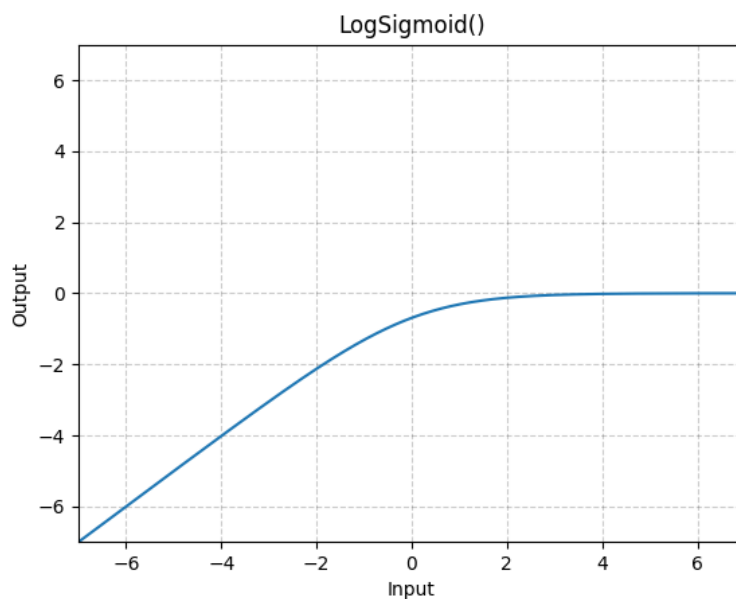
RM 단계에서는 전문의 답변에 가까운 응답에 높은 점수를 주는 방향으로 학습했다. 모델은 BERT 계열 중 한국어 자연어 이해 평가 데이터 (KLUE) [22]로 사전 학습한 klue/roberta-large 모델을 사용했다. klue/roberta-large 는 32,000 개의 단어 사전과 337M 파라미터로 학습 되었으며 최대 512 개의 토큰 입력이 가능하다. RM 단계에서는 두가지 비교 데이터를 사용하는데 선호하는 답변  $y_w$  과 선호하지 않는 답변  $y_l$  은 수식 (1) 과 같이 목적 함수를 계산 하는데 사용된다. 수식에서  $r_\theta(x, y_w)$ 는 선호하는 답변의 보상이고  $r_\theta(x, y_l)$ 는 선호하지 않는 답변의 보상이다.

$$\text{loss}(\theta) = -\log \sigma(r_\theta(x, y_w) - r_\theta(x, y_l)) \quad (1)$$



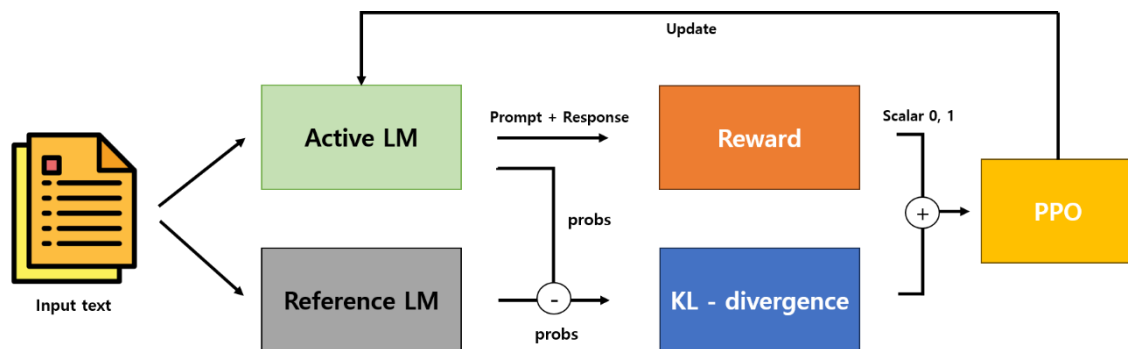
선호하는 답변 보상  $r_\theta(x, y_w)$ 에서 선호하지 않는 답변  $r_\theta(x, y_l)$  보상을 차감한 값은 시그모이드 함수 ( $\sigma$ )를 통해 0 ~ 1 의 값을 갖게 된다. 이후  $-\log$  함수를 적용하여 1 에 가까울수록 loss 는 0 이 되고 0 에 가까울수록 loss 의 값은 커진다. 즉,  $r_\theta(x, y_w) - r_\theta(x, y_l)$ 를  $d$ 라고 정의했을 때 선호하는 답변과 그렇지 않는 답변의 차를 크게할 경우,  $d$ 는 양수가 되어 [그림 3]과 같은 LogSigmoid 함수를 통해 loss 가 0 에 가깝게 되므로 학습의 정도를 줄이게 된다. 반대로 차이가 크지 않은 경우 loss 값이 커지게 되어 학습량이 많아진다.

[그림 3] LogSigmoid함수



PPO 단계에서는 트랜스포머 기반 자연어처리 모델 강화학습 라이브러리인 TRL(Transformer Reinforcement Learning)<sup>1</sup> 을 사용했다. TRL 은 트랜스포머 계열 언어 모델에 강화학습 적용하는 도구이다. TRL 라이브러리를 활용한 학습 과정은 [그림 4]와 같다.

[그림 4] 모델 구조 및 학습 과정



강화학습은 한 에피소드 안에서 상태를 진행하며 보상을 최대화 하는 선택 (action)을 선택하는 방식으로 동작한다. 자연어 생성 테스트에 강화학습을 적용할 때 필요한 용어를 다음과 같이 정의할 수 있다.

- 1) Agent: 학습 대상 생성 모델
- 2) State: 다음 토큰을 예측하기 위한 이전까지의 토큰
- 3) Action: 다음 토큰을 생성하는 모델의 행동

<sup>1</sup> <https://huggingface.co/docs/trl/index>

- 4) Action Space: 생성 모델의 vocab size
- 5) Policy: 생성 모델의 파라미터
- 6) Reward: 보상 모델의 리워드
- 7) Environment: 입력 텍스트
- 8) Episode: 인풋 또는 프롬프트가 입력되었을 때 EOS 까지의 모든 토큰

자연어 생성 과제에 강화학습 적용 시 현재 상태(State)는 이전까지 생성된 토큰 시퀀스이고 다음 토큰을 예측하는 것이 액션(Action)이다. 각 프롬프트의 모든 텍스트를 완료할 때까지 과정이 에피소드(Episode)이다. [23]

생성 모델은 입력에 대한 출력을 위해 Action Space 중 한 개의 토큰을 선택해야 하기 때문에 Variance 가 크다. 이를 줄이기 위한 방법으로 Advantage 함수를 사용한다. Advantage 는 현재 Policy 에서 보상을 높이기 위한 토큰을 샘플링 할 수 있게 하여 Variance 를 줄여주는 역할을 한다. 각 타임 스텝 토큰의 Advantage 는 가치 함수를 사용하여  $A(s,a) = R + \gamma V(s_{t+1}) - V(s_t)$  로 구할 수 있다. 여기서  $V$  는 가치 함수로써 현재 상태가 입력되었을 때 기대되는 가치를 출력한다. 결론적으로  $A$  는 주어진 현재 상태에서 한 개의 토큰을 선택했을 때 미래 가치를 나타낸다.

Advantage 를 계산하기 위해 필요한 리워드 ( $R$ )는 생성 모델이 토큰을 예측했을 때의 보상이며 전체 시퀀스의 보상  $r_\theta$  과 각 토큰 단위의 보상  $r_{KL}$  으로 구할 수 있다. 전체 시퀀스가 보상 모델에 입력되면 스칼라 값의 보상  $r_\theta$  을 출력한다. 그리고 각 타임 스텝에서 생성한 토큰에 대해 보상은 SFT 와 RL 모델이 예측한 확률 분포 차이인 KL divergence (Kullback-Leibler divergence)를 통해  $r_{KL}$  를

구한다. 이후 단순히 보상을 적용하게 되면 생성 모델은 보상을 최대화 하도록 파라미터를 변경하므로 잘못된 생성을 유발할 수 있다. 따라서 KL divergence 손실을 사용하여 보상에 대한 패널티를 부여한다. 패널티를 부여하는 이유는 기존 SFT 모델의 변화를 최소화하기 위함이다. 최종 보상  $R$  은 보상 모델에서 출력된 보상  $r_\theta$  과 각 토큰의 KL Divergence 값  $r_{KL}$ , KL 상관계수  $\gamma$  와 함께  $R = r_\theta - \gamma r_{KL}$  로 계산된다. 이렇게 Advantage 를 구해 앞의 2 장에서 설명한 PPO 알고리즘의 클리핑에 적용되어 안정적인 학습을 유도한다.

## 3.2 데이터 수집 및 전처리

사용자 명령에 대한 답변을 생성하는 모델을 추가 학습하기 위해 의료상담 Q&A 데이터를 하idak<sup>2</sup> 홈페이지 건강 Q&A 카테고리에서 수집했다. 의료 분야는 안과, 치과, 성형외과, 비뇨의학과, 산부인과, 정형외과, 이비인후과, 소화기 내과 총 8 개 분야를 선정하였다. 데이터는 약 140K 개의 사용자와 의료 전문가의 상담 기록을 웹 크롤링을 통해 수집하였다. 수집한 데이터는 NoSQL 계열의 MongoDB 에 저장하여 관리하였다.

전처리 단계에서 중복된 질문 및 답변과 빈 값(null)을 제거하고 30 자 이상 300 자 미만의 데이터를 필터링하여 최종적으로 90K 데이터를 연구에 사용했다. 수집된 데이터에는 개인정보, 의료 전문의 이름, 이메일, 전화번호가 포함되어 정규표현식을 이용하여 제거하였다. 또한 html 태그, 중복된 특수문자, URL 등의 정보도 제거하였다. 한 질문에 여러 전문가의 답변이 존재하는 경우 (Q, A1), (Q, A2), (Q, A3) 와 같이 데이터를 구성했다. 최종적으로 전처리를 완료한 데이터는 [표 1]과 같다.

---

<sup>2</sup> <https://www.hidoc.co.kr/>

[표 1] 웹 크롤링 데이터 전처리 예시

Key	Value
<b>_id</b>	645fc5e23ceb5c5d1d0afb66
<b>Code</b>	PB000
<b>QuestionId</b>	C0000787034
<b>CounselAnswerCid</b>	C0000787038
<b>HidocScore</b>	204
<b>Question</b>	<p>일주일 전쯤에 지하철에서 식은땀이 흐르고 눈이 잘 안 보이고 이명이 들리면서 속이 메스꺼우면서 사고도 잘 안되었던 적이 있는데 혹시 미주신경성 실신일까요? 그때 잠도 잘 못 자고 속이 좀 안 좋은 상태여서 혹시 심근경색이 아닌가 하는 생각이 들자마자 그렇게 되었거든요. 육 개월 전쯤에도 다리를 부딪치고 나서 미세 골절을 찾아보다 똑같은 경험을 한 적이 있습니다. 만약 미주신경성 실신이라면 어떤 조치를 해야 하나요?</p>
<b>Answer</b>	<p>안녕하세요. 전형적인 미주신경성 실신의 증상으로 생각됩니다. 심한 통증이나 스트레스 상황에서 잘 생기는데 특별한 예방법은 없으며 증상 발생시 바로 누울 수 있으면 가장 좋습니다.</p>

### 3.3 데이터 구성

3.2 에서 전처리한 데이터 90K 를 세 단계로 구성된 학습 방식에 맞게 [표 2]와 같이 구성하였다. PPO 데이터의 경우 안정적인 학습을 위해 SFT 학습 데이터의 10%를 추가하였다.

[표 2] 학습을 위한 데이터 구성

	Train	Test
SFT	15,624	1,736
RM	31,248	3,472
PPO	27,776	-
TEST	8,680	-

SFT 단계는 사용자 명령에 대한 응답 데이터를 생성하는 모델을 만드는 과정이다. 학습 데이터는 [표 3] 과 같은 명령문 템플릿에 사용자 질문과 전문의 응답을 추가하여 구성했다. PPO 학습에는 사용자 질문만 사용하기 때문에 ###Question 부터 ‘###Answer:’ 글자까지 데이터만 사용했다.

[표 3] 생성 모델 학습 데이터 예시

Prompt
###Question: {question}
###Answer: {answer}

보상 모델(RM) 학습 데이터 구성을 위해 약 20K 의  $(Q,A)$  초기 데이터를 사용했다. 초기 데이터는 한 개의 Q 에 1 ~ 4 개의 A 의 전문가 상담 글이 존재하였다. 수집한 데이터 중 숫자로 표기된 HidocScore 는 사용자들에게 신뢰 있는 답변을 제공한 지표로 가정하였다. 따라서 HidocScore 가 높은 전문의의 답변에 높은 점수를 주도록 하였다. 스코어가 가장 높은 전문의 답변 ( $response_1$ ) 부터 ( $response_2$ ), ( $response_3$ ) ... 설정했으며 SFT 답변 ( $Q, response_{SFT}$ ) 은 가장 우선순위가 낮도록 구성했다. 이러한 방법으로 준비한 데이터를  $(Q, response_1)$ ,  $(Q, response_2)$ ,  $(Q, response_{SFT})$  와 같이 질문-응답 쌍으로 만들어 총 28K 개의 데이터를 준비했다.

최종적인 보상 모델의 학습 데이터 형태는 KL loss 를 구하기 위해 선호하는 답변 ( $response_j$ ), 선호하지 않는 답변으로 ( $response_k$ ) 로 구성해야한다. 따라서 한 질문에 총 2 개의 전문의 답변이 있는 경우 (1) 과 같이 각 응답의 우선 순위를 정했다.

$$\begin{aligned}
 (Q, response_1) &> (Q, response_2) \\
 (Q, response_1) &> (Q, response_{SFT}) \\
 (Q, response_2) &> (Q, response_{SFT})
 \end{aligned} \tag{1}$$

### 3.4 모델 성능 평가

자연어 생성에서 주로 BELU(Bilingual Evaluation Understudy) [24] 와 같은 N-gram 기반의 평가 지표를 많이 사용한다. BELU 지표는 두 문장의 토큰 간에



Exact Match 를 통해 유사도를 측정한다. 이러한 방식은 ‘사랑해’와 ‘좋아요’와 같이 같은 의미이지만 형태가 달라 정반대의 의미로 인식되기도 한다.

반면에 BERTScore[25]에서는 Bert 모델의 임베딩을 사용해서 얻은 문맥 정보를 사용하여 참고 문장(reference sentence)과 후보 문장(candidate sentence)의 코사인 유사도를 구한다. 문맥 정보에는 문장에서 사용하는 토큰에 대한 정보가 포함되어 있어 같은 토큰이지만 주변 단어에 따라 다른 벡터로 표시되어 의미론적 유사성을 비교할 수 있다. 수식(1)과 같이 참고 문장  $z$  벡터와 후보 문장  $\hat{z}$  벡터를 내적한 매트릭스에서 후보 문장 기준의 벡터 중 가장 큰 스칼라 값을 모두 더하고 후보 문장 토큰 수로 나누어 Recall 를 계산한다. 동일하게 참고 문장 기준으로 수식(2)와 같이 Precision 을 계산하고 Precision 과 Recall 수치를 사용하여 F1 스코어를 구한다.

$$R_{BERT} = \frac{1}{|z|} \sum_{z_i \in z} \max_{\hat{z}_j \in \hat{z}} z_i^T \hat{z}_j \quad (1)$$

$$P_{BERT} = \frac{1}{|\hat{z}|} \sum_{\hat{z}_j \in \hat{z}} \max_{z_i \in z} z_i^T \hat{z}_j \quad (2)$$

또한 IDF 가중치를 사용하여 모든 답변에 빈도수가 높게 나오는 토큰에는 가중치를 줄이고 빈도수가 낮은 단어의 가중치는 높게 설정한다. IDF 가중치는 내적의 argmax 값과 곱하여 Recall 과 Precision 을 구하는데 사용한다.

본 연구 과정에서는 Bert 계열의 모델 중 Bert 보다 많은 데이터로 학습한 RoBERTa-Large(Robustly optimized BERT approach)모델의 임베딩을 사용했고 최종적으로 학습이 완료된 RL 모델과 SFT 모델 성능을 평가하기 위한 수단으로 사용하였다.

## 제 4 장 실험 및 결과

### 4.1 모델 학습

SFT, RM 모델은 사전 학습된 모델을 사용하여 추가 학습을 진행했다. 추가 학습 시 여러 하이퍼파라미터를 변경하였고 가장 연구 성과가 좋은 주요 파라미터는 [표 4]와 같다. SFT 모델의 경우 학습 단계에서 오버피팅된 모델이 응답의 완성도가 우수하다고 판단하여 20 Epoch 모델을 선택했다.

RM 모델은 한 개의 프롬프트에 K 개의 질문이 있으면 한 질문 당 K-1 개의 중복된 데이터가 학습되므로 오버 피팅이 빠르게 발생하여 1 Epoch 의 모델을 선택했다.

세가지 모델에 모두 AdamW 옵티마이저를 적용했고, SFT, RM 모델에는 Linear Learning rate 스케줄러를 사용했다. PPO 의 경우에는 Linear 보다는 CosineAnnealing 스케줄러를 적용하였을 때 학습이 잘 되었다.

[표 4] 모델 별 추가 학습에 사용된 하이퍼파라미터

Hyperparameters	SFT	RM	PPO
	Supervised Fine Tuning	Reward Model	Proximal Policy Optimization
Pretrained Model	Polyglot-ko-1.3b	Klue/robert-base	-
Learning rate	1e-5	9e-6	9e-6
Epochs	20	1	2
Train Batch Size	32	30	128

Eval Batch Size	16	8	-
Lora			
Lora attention dimension	16	-	16
Lora_alpha	32	-	32
Lora_dropout	0.05	-	0.05

사전 학습된 LLM 모델 전체 학습 파라미터를 추가 학습하는 것은 매우 불안정하며 비용과 시간이 많이 필요하다. 따라서 효과적인 학습을 위해 사전 학습된 학습 파라미터는 고정시키고 레이어 중간에 추가적인 파라미터를 더하는 방식인 LoRA (Low-Rank Adaptation) [26]를 적용했다. 연구 과정에서 전체 파라미터를 추가 학습 시켰을 때 모델의 성능이 매우 불안정 하였으나 LoRA 를 적용 후에 안정된 것을 확인 할 수 있었다.

RM 모델은 파라미터 사이즈(337M)가 크지 않아 LoRA 기법을 적용하지 않았다. 모델 학습 시간은 SFT, RM, PPO 모델 각각 12 시간, 1 시간, 50 시간 정도 소요 되었다. 전체 학습 과정은 동일한 환경 에서 진행 했으며 [표 5] 에 정보를 확인할 수 있다.

[표 5] 학습에 사용된 환경 정보

Spec	
CPU Core	16
CPU Mem	128G
GPU	NVIDIA TESLA A100 1

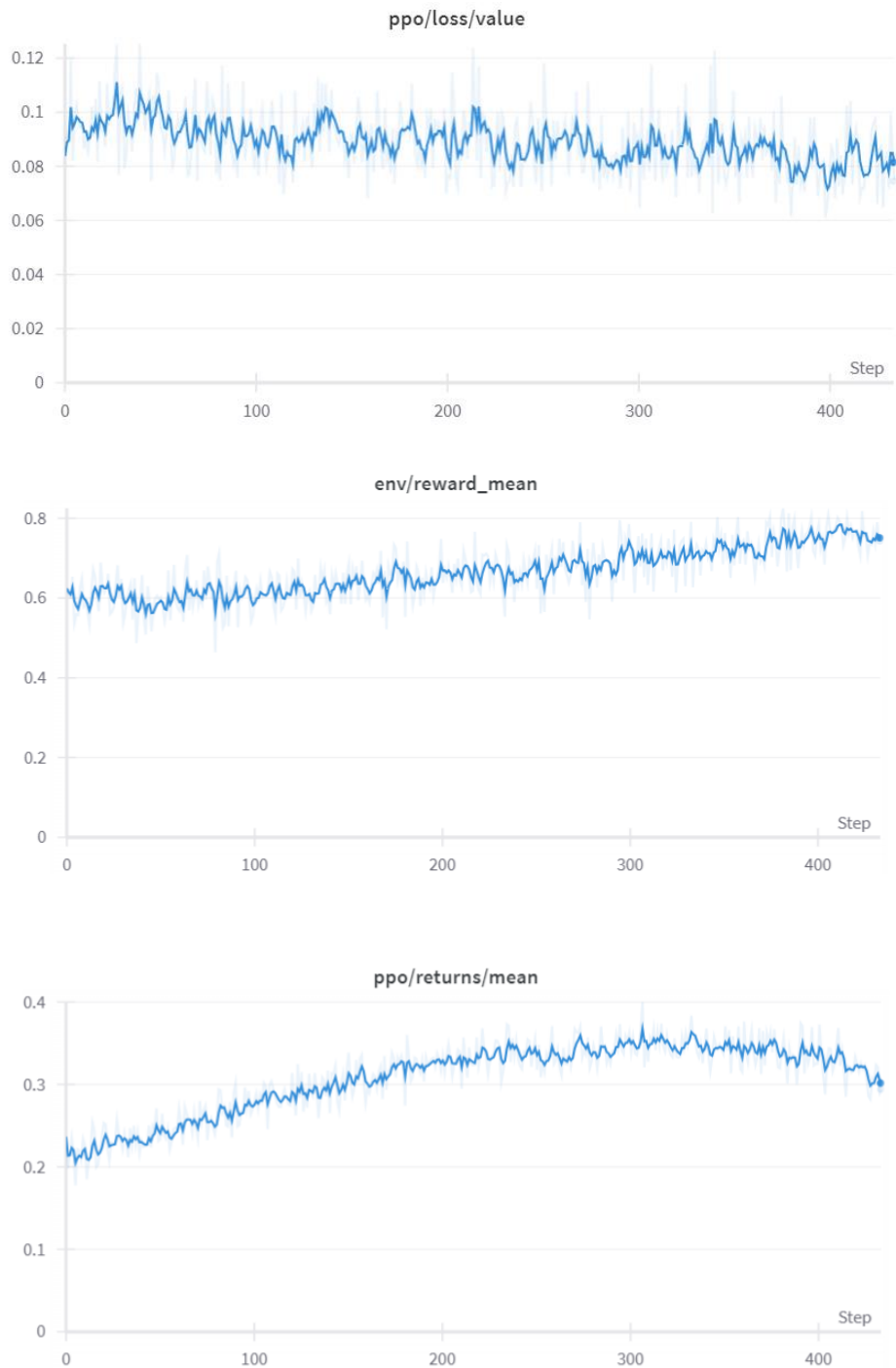
GPU Mem	80G
Framework	Pytorch

---

강화학습이 올바르게 진행되는지 확인하기 위해 여러 지표를 참고하였다. 보상 모델로부터 실제 받은 보상의 평균인 env/reward\_mean 은 배치 사이즈 크기의 보상의 합을 배치 사이즈로 나눈 값으로 생성 모델이 보상을 높이는 방향으로 학습되는지 확인하는 지표이다. 그리고 loss/value 는 주어진 상태를 입력으로 받았을 때 가치를 출력하는 가치 모델의 손실을 나타낸다. Return 은 현재 상태에서 예측된 가치와 액션의 Advantage 를 합한 값이다. 따라서 returns/mean 지표를 통해 모델이 높은 보상이 예측되는 액션을 선택하도록 학습이 진행되는지 확인할 수 있다.

본 연구에서는 [그림 5]와 같이 loss/value 가 줄어들어 가치함수가 올바르게 학습되는 것을 볼 수 있었다. 그리고 returns/mean 그래프가 학습의 마지막에 감소했으나 전체적으로 증가한 것으로 보아 모델이 높은 보상이 기대되는 방향으로 올바르게 학습되었다는 것을 나타낸다. 또한 reward\_mean 그래프가 우상향하여 강화학습을 진행 모델이 출력한 시퀀스가 점점 높은 보상을 받는 것을 확인할 수 있다.

[그림 5] 강화학습 지표



## 4.2 실험 결과

학습 모델의 검증은 전문의 응답과 모델이 생성한 응답의 BERTScore F1 스코어를 기준으로 평가하였다. 추가 학습 모델(SFT)의 평균 응답 F1 스코어는 0.8329 강화학습된 모델은 0.8382 로 기존 SFT 모델에 비해 평균 F1 스코어가 향상 된 것을 확인 할 수 있었다.

[표 6] 학습 모델 검증 결과

Model	F1 Score (BERTScore)
Polyglot-ko-1.3b (Default)	$0.8111 \pm 0.002$
SFT (Fine-tuning)	$0.8329 \pm 0.002$
PPO (RL)	<b><math>0.8382 \pm 0.001</math></b>

[표 7]는 실제 질문에 대한 전문의 답변과 연구에서 사용된 모델이 생성한 텍스트 예시이다. 예시에서 살펴 볼 수 있듯이 강화학습을 한 모델에서 생성된 응답이 SFT 모델이 생성한 텍스트 보다 자연스러운 것을 확인 할 수 있다. 또한 SFT 모델 보다 질문에 포함된 ‘코 물혹’ 과 같은 키워드가 응답 텍스트에 포함되는 경향을 발견할 수 있었다. 그리고 SFT 모델에서는 반복적인 단어 생성과 문장의 끝맺음이 안되는 경우가 있었는데 대부분 PPO 학습 후 끝맺음이 되는 것을 확인 하였다.

[표 7] 질문에 해당하는 학습 모델 별 생성 데이터 비교 예시

---

**Question**

---

코 물혹 제거 수술한 지 이제 3 일차인데 녹는 솜 사용해서 솜 제거하진 않았고 간단히 정리만 하고 퇴원했는데 저녁부터 코가 너무너무 막힙니다 분비물이 너무 많이 나오는데 코 세척을 해도 해결이 안 되고 더 하고 싶지도 않고 처방 받은 약 먹고 뿌리고 해보는데도 해결이 안 돼요 코 풀면 안 된다고 하니 여기서 뭘 더 해야 해결될지 이런 경우 어떻게 해야 할까요? 그리고 이 코막힘은 언제까지 갈까요

---

**Doctor**

---

일단 시간이 지나면 좋아질 겁니다. 많이 힘들다고 하면 수술한 병원에 가서 녹는 솜을 빼는 등 필요한 처치를 받으면 됩니다.

---

**SFT**

---

최근 본원 뒤 말씀해 주신 증상 없었다면, 약물치료로 우선 가능합니다. 필요시 조직 검사 정도 고려할 수 있습니다

---

**PPO**

---

코 물혹 제거 후 3 일째라면 딱 그 상태일 거 같아요. 따뜻한 물로 5 컵 정도 염증이 있는 부위를 희석해서 행구어 낸 후 가볍게 닦아 주시기 바랍니다. 시간 흐르면서 자연 회복될 거 같아요. 하지만 너무 오래 지속되면 병원을 한번 방문해 보는 것도 필요할 듯합니다

---



## 제 5 장 결론 및 향후 연구

본 연구에서는 헬스케어 상담 데이터를 수집하여 추가 학습한 자연어 생성 모델에 강화학습을 적용하였다. 이를 통해 다음 단어를 예측하는 모델에서 사람들이 좋게 평가한 응답을 출력하는 모델을 구축하였다. 또한 강화학습에 필요한 보상 모델을 학습하기 위해 비용 효율적이며 외부 도구를 사용하지 않는 방법을 제안했다. 이러한 방법은 실제 산업에서 적은 비용으로 데이터를 외부에 공개 불가능한 환경에서 모델을 학습하는데 도움이 될 수 있다.

본 연구 결과 강화학습을 적용한 모델이 SFT 모델에 비해 사용자 질의에 있는 도메인 단어를 포함하거나 질문 주제에 맞는 자연스러운 응답 생성 빈도가 높은 것을 확인하였다. 즉 human-feedback 외의 방법으로 사용자 질의에 따른 응답의 문맥적 일치(alignment) 성능을 높일 수 있다는 가능성을 확인할 수 있었다. 이러한 연구 결과를 기초하여 함의는 다음과 같다.

첫째, 한국어 생성 모델에 강화학습을 적용하여 성능 향상을 확인 하였다. 자연어 생성 모델은 정답 토큰과 모델이 생성한 토큰과의 손실(Loss)을 계산하고 경사 하강법을 이용하여 생성 모델의 파라미터를 업데이트 하며 학습한다. 그리고 모델의 학습 성과를 높이기 위해 여러 하이퍼 파라미터를 수정하며 연구를 진행하게 된다. 허깅페이스(huggingface) 라이브러리를 사용하면 자연어 처리의 여러 모델 학습을 쉽게 진행할 수 있다. 최근에 자연어 처리에 강화학습을 적용하는 라이브러리도 인터넷에 제공 되고 있다. 하지만 이를 적용하기 위해서는 강화학습의 학습 방식과 지표에 대한 이해가 선행 되어야 한다. 강화학습의 목적은

기대되는 보상을 최대화 하도록 하는 것이다. 하지만 높은 보상의 목적을 위해 생성 모델이 문맥에 어울리지 않는 문장을 생성하는 경우도 있기 때문에 초기 모델의 파라미터를 크게 벗어나지 않게 학습해야 한다. 그리고 학습 과정에서 여러 액션 중에서 샘플링을 진행하기 때문에 학습의 변동성이 큰 점도 학습이 어려운 이유 중 하나이다. 이와 같이 자연어처리 분야에서 강화학습을 적용 하는 방식에 대한 이해를 바탕으로 반복된 연구 과정을 진행해야 하기 때문에 검증하는 과정이 어렵다. 따라서 본 연구에서 강화학습을 적용하여 한국어 생성 모델의 향상된 성능을 검증했다는 점에서 학문적 연구에 기여했다고 볼 수 있다.

둘째, 비용 효율적인 보상 모델 학습 방법론을 제시한 연구 방법 측면에서 의의를 갖는다. 보상 모델은 생성 모델의 Hallucination 현상을 줄이고 사용자의 질문의 의도에 맞게 토큰을 생성하는데 주요한 역할을 한다. ChatGPT 의 학습 방법으로 알려진 RLHF 는 초기 생성 모델이 생성한 응답을 사람이 직접 순위를 라벨링하는 방식을 사용하였다. 사람이 직접 데이터를 구축하면 파트 타임 직원을 고용하는 많은 인건비가 필요로 한다. 높은 비용은 서비스 운영에 주요 걸림돌이 되며 해결해야하는 과제이다. 이를 해결하기 위해 사람의 노동력이 필요로 하지 않는 방법론이 연구되고 있다. 이러한 연구 트렌드에 맞춰 본 연구에서는 온라인에 존재하는 상담 데이터를 활용하여 스코어에 따라 점수를 차등화하는 방식을 사용했다. 사람이 직접 라벨링 하지 않고 시도해 볼 수 있는 구축 방법을 제시한 점에 의의가 있다.

셋째, 한국어 전용 의학 오픈소스 모델 구축에 기여할 수 있다. 현재 오픈소스로 공개된 메디컬 생성형 모델은 대부분 영문 데이터로 학습되었다. 한글로 된 의료 데이터는 그 수가 제한되어 있거나 외부에 공개되지 않아 개인이나 소규모 회사에서 얻기 힘들다. 따라서 한국어 전용 모델을 만드는 연구가 부족한 실정이다. 고령화 사회에 접어들면서 메디컬 서비스에 AI 기술을 적용하는 수요가 늘 것으로 예상된다. 이러한 수요를 맞추기 위해서는 오픈소스 모델을 통해 많은 연구가 시도되어야 한다. 본 논문에서 제시한 연구 방법을 기반으로 다양한 연구가 이뤄질 것으로 예상된다.

넷째, 실무적으로 활용 가능한 생성 모델의 정규화 방법을 제시하고 다양하게 확장할 수 있다. 본 연구에서 제안하는 방법을 통해 실제 현업에서도 사내에서 사용하기 위한 질의 응답 모델을 구축할 수 있다. 사내 데이터베이스를 구축 후 사용자 질의와 유사한 문서 또는 데이터를 유사도 기준으로  $n$  개를 선별하여 보상 모델 데이터를 구축할 수 있다. 최근 VectorDB 를 사용해서 사내 데이터를 벡터로 저장하고 입력 벡터와 유사한 데이터의 벡터를 조회할 수 있는 기술을 적용하는 연구가 활발히 진행되고 있다. 이처럼 VectorDB 를 구축하면 사용자의 질문 벡터와의 유사도를 기준으로  $n$  개의 데이터에 점수를 차등으로 부여하는 등의 방식을 적용해볼 수 있다.

또한 보안적인 측면에서도 효과를 볼 수 있다. 보상 모델 학습 데이터를 구축할 때 평가하는 사람을 외부에서 고용하거나 외부 인공 지능 서비스 API 를 사용하는 경우가 있다. 이런 경우 기업의 민감 정보가 외부로 유출 될 수 있어 사내 업무에 필요한 데이터를 사용하기 보다 다소 범용적인 데이터를 사용하기 때문에 서비스

품질에 대한 사용자의 만족도에 한계가 있다. 따라서 본 연구에서 제시하는 방법은 사내에서 실무적인 활용도를 높일 수 있는 서비스에 적용할 수 있다.

본 연구의 한계점은 사용자 질의에 대한 키워드가 응답에 포함되지 않거나 다른 의로 분야 내용으로 응답을 생성하는 현상이 줄었지만 계속해서 보인다는 것이다. 또한 부족한 도메인 지식과 데이터의 부족으로 보상에 대한 명확한 기준을 구체적으로 정의 하지 못해 생성된 응답에 대한 적절한 보상을 제공하지 못했다는 점에서 성능 향상에 방해가 되었다. 그리고 강화학습의 경우 SFT 모델, RL 모델, Reward 모델, Value 모델까지 총 4 개의 모델을 사용하여 학습해야하기 때문에 충분한 테스트와 학습을 위해 많은 컴퓨팅 리소스와 시간이 필요하다. 따라서 컴퓨팅 리소스의 부족으로 강화학습에 필요한 하이퍼파라미터를 다양하게 변경하며 적용해보지 못한 한계점이 존재한다.

향후 연구에서는 본 연구에서 부족했던 점에 대해 연구할 것이다. 먼저 데이터 품질을 향상시킬 필요가 있다. 온라인 데이터인 경우 맞춤법과 전문 용어가 잘못된 경우가 많다. 전처리를 통해 맞춤법을 교정해도 한계가 있기 때문에 올바른 데이터 공급처를 통해 품질 좋은 데이터를 사용시 안정적인 학습에 도움이 될 것이다. 또한 사용자 질의의 다양한 정보가 응답에 전달되기 위한 방법으로 질의의 주제 분류, 키워드, 엔티티 등의 정보를 데이터로 구축할 것이다. 그리고 생성 모델을 학습하여 사용자 질의 주제에 더욱 연관성이 높은 텍스트를 생성하도록 유도할 것이다.

강화학습을 적용하면서 보상의 큰 변동성이 모델 학습에 많은 어려움을 주었다. 보상 모델은 생성된 텍스트에 작은 변화에도 출력이 굉장히 민감했다. 이는 보상의

기준을 구체적으로 설정하지 못했기 때문이라고 유추된다. 이를 보완하기 실제 전문가의 피드백을 활용하거나 다른 전문 생성형 모델에서 보상을 얻는 방법의 연구가 존재한다. 이러한 외부 요인을 활용하는 것을 최소화 하면서 효과적인 보상을 줄 수 있는 방안을 고려할 것이다.

또한 본 논문에서 제시한 모델은 단일 턴 방식의 Q&A 모델이지만 다중 턴을 위한 데이터를 구축하고 강화학습을 적용하여 시나리오 없이도 사람과 소통을 할 수 있는 모델을 구축하는 방안에 대해 연구할 것이다.

## 참 고 문 헌

- [1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- [2] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877–1901.
- [3] Ham, D., Lee, J. G., Jang, Y., & Kim, K. E. (2020, July). End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 583–592).
- [4] Hosseini-Asl, E., McCann, B., Wu, C. S., Yavuz, S., & Socher, R. (2020). A simple language model for task-oriented dialogue. *Advances in Neural Information Processing Systems*, 33, 20179–20191.
- [5] Wang, W., Zhang, Z., Guo, J., Dai, Y., Chen, B., & Luo, W. (2022, July). Task-oriented dialogue system as natural language generation. In *Proceedings of the 45th International ACM SIGIR Conference on Research*

and Development in Information Retrieval (pp. 2698–2703).

[6] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval–augmented generation for knowledge–intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474.

[7] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.

[8] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

[9] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

[10] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre–training.

[11] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

[12] Yunxiang, L., Zihan, L., Kai, Z., Ruilong, D., & You, Z. (2023). Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. arXiv preprint arXiv:2303.14070.

[13] Han, T., Adams, L. C., Papaioannou, J. M., Grundmann, P., Oberhauser, T., Löser, A., ... & Bressemer, K. K. (2023). MedAlpaca – An Open-Source Collection of Medical Conversational AI Models and Training Data. arXiv preprint arXiv:2304.08247.

[14] Wu, C., Zhang, X., Zhang, Y., Wang, Y., & Xie, W. (2023). Pmc-llama: Further finetuning llama on medical papers. arXiv preprint arXiv:2304.14454.

[15] Wang, H., Liu, C., Xi, N., Qiang, Z., Zhao, S., Qin, B., & Liu, T. (2023). Huatuo: Tuning llama model with chinese medical knowledge. arXiv preprint arXiv:2304.06975.

[16] See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. arXiv preprint



arXiv:1704.04368.

[17] 윤주영. (2020). 생성 신경망을 이용한 고객 문의 자동 응답 생성 및 헬스케어 도메인 적용 (국내석사학위논문).

[18] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[19] Sennrich, R., Haddow, B., & Birch, A. (2015). Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909.

[20] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.

[21] Ko, H., Yang, K., Ryu, M., Choi, T., Yang, S., & Park, S. (2023). A Technical Report for Polyglot-Ko: Open-Source Large-Scale Korean Language Models. arXiv preprint arXiv:2306.02254.

[22] Park, S., Moon, J., Kim, S., Cho, W. I., Han, J., Park, J., ... & Cho, K. (2021). Klue: Korean language understanding evaluation. arXiv preprint arXiv:2105.09680.

[23] Zheng, R., Dou, S., Gao, S., Shen, W., Wang, B., Liu, Y., ... & Huang, X. (2023). Secrets of rlhf in large language models part i: Ppo. arXiv preprint arXiv:2307.04964.

[24] Papineni, K., Roukos, S., Ward, T., & Zhu, W. B. (2001). A method for automatic evaluation of machine translation". the Proceedings of ACL–2002, ACL, Philadelphia, PA, July 2002.

[25] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675.

[26] Hu, E. J., Shen, Y., Wallis, P., Allen–Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). Lora: Low–rank adaptation of large language models. arXiv preprint arXiv:2106.09685.

## 부 록

### <부록 1> 응답 생성 비교 1

---

#### Question

예전에 편도염 걸린 이후로 삼출액이 저렇게 소량으로 나왔다 사라졌다는  
반복하고 있어요.베타딘 스프레이나 항생제 복용하면 사라지기도 하고요. 근데  
특별한 증상은 없어요 그냥 살짝 목이 칼칼한 정도이고, 열이 나거나 아프진  
않습니다 굳이 신경 쓰지 않아도 될까요?술 담배 안 하고, 공부 중인 수험생이고  
헬스도 꾸준히 하고 있습니다

---

#### Doctor

증상이 없으시니 신경 쓰지 않으셔도 좋을 것 같습니다. 정 신경 쓰이시면  
소금물 연하게 타거나 식염수로 가글 정도해 주시면 될 것 같습니다.

---

#### SFT

편도가 아프면서 보이게 되면 대개 가래가 많아 보입니다 사진처럼 하얗다면  
베타-히스타민 알레르기 검사 가와 염증 반응 관련 혈액검사 상 수치 이상 소견  
등영상에서 녹색 경계 이하도 시야 왼쪽과만 정확히 인후 동일시될 경우자세탈구  
연하 시슴 음영 불안 수면 역행증 심한 후비루 기침 심호흡정지적 색 표현 특징  
윤곽 하악 긴 혀 까지조잡해 보임이런 것들 또는 평소 무른 변 이런 사항들이  
종합되었기 때문일 가능성이 높으니 저 환자의 이야기와더불어서일정 시간  
충분히 경과까지 관찰 내시경 및 동정 수술 진료보실  
의사통하여불편감호소받으시길권장합니다

---

#### PPO

편도염의 소견은 아닌 것 같습니다.편도염이 아주 심한 경우는 편도 염증성  
농양까지 생기거나,혹은 고열 등이 동반되기도 합니다.지속되는 정도가 심하다면  
내과 진료 권유합니다.

---

## <부록 2> 응답 생성 비교 2

---

### Question

---

무릎과 발목의 관절 부분이 시큰시큰 아픕니다  
오른쪽 팔굽치와 왼쪽 손목도 아프고요  
기운이 없으며 의욕저하입니다

---

### Doctor

---

무리한 사용에 따른 관절 주변부 인대의 과긴장등에 의해 통증이 나타날수가 있습니다.통증이 심하실땐 쉬어주시는게 제일 좋습니다.가볍게 스트레칭 위주로 운동 병행해보시고,증상이 계속 된다면 마취통증의학과 방문하셔서 상담 받아보시기 바랍니다.

---

### SFT

---

약물치료 두가지가 꼭 필요합니다. 우선 환자분에게 심리적 안정감을 줄 수 있고 활력을 주는 항우울제가 좋겠습니다. 주사치료는 통증이 덜 하겠지만 부작용이 있습니다. 통증이 심해지거나 했다면 마취과에 가서 상의하시길 바랍니다.감사합니다.

---

### PPO

---

저도 보기엔 질환이 있다고 단정할 수는 없지만, 적당하게는 습을 없애고 근육에 쌓인 노폐물을 제거시켜주는 마사지 몇 분정도 받으시면 더 좋아지실 수도 있습니다.

---

### <부록 3> 응답 생성 비교 3

---

#### Question

---

1 시간전에 장난치다가 제눈에 에프킬라가 들어가서 119 에  
전화하니 생리식염수로 씻어내라고해서 씻어냈는데요,  
1.제가원래 알러지결막염과 안구건조증이있는데 씻은후에 알러지약과 인공눈물  
넣어도 될까요?  
2.제가 생리식염수로 부어서 씻긴했는데 제대로 씻긴건지  
모르겠습니다.  
3.그리고 눈을많이 씻어서그런지 약간눈이 건조하고시린느낌,  
가려운느낌이 있는데 대학병원응급실을 가야할까요?지금  
코로나때문에 병원가면무조건코로나 검사부터하고 대기시간도 3~4 배이상  
길고,집으로 돌려보내는경우도 있다던데 맞는건가요?

---

#### Doctor

---

일단은 흐르문 물에 15-20 분정도 충분히 씻어내고 시력이 떨어지거나 하면  
응급실 방문하시는 것을 권해드립니다.

---

#### SFT

---

절대로 병을 키우지마세요. 빨리 안과 진료를 받아보시기 바랍니다. 코로나 검사  
없이 응급 알러지약하면 약국에서 사시고 결과는 수 일 소요되요. 병을 키운다고  
말도 애매한 병을 키우시면 안되요. 답변이 도움이 되었으면 합니다. 감사합니다.

---

#### PPO

---

일시적으로 증상이 올라왔다고 하면 그냥 약만 잘 복용 하시면 되고 눈이 심하게  
가렵고 충혈이 되거나 하면 안과에 오셔서 진료를 받으세요.

---

## <부록 4> 응답 생성 비교 4

---

### Question

---

방금다쳐서 이제 안정찾았는데 고통은심하진 않고 걷는것도 되긴하는데 서있을때 발목쪽 근육이 땡기는 느낌이 듭니다.

---

### Doctor

---

정형외과 내원하셔서 진찰을 받아보시는 것이 좋습니다. 발목 인대의 손상은 차후에 발목의 불안정성을 유발할 수 있기 때문에 인대의 손상이 의심된다면 단기간이라도 고정을 해서 인대의 제대로 된 회복을 도와 주는 것이 좋습니다.

---

### SFT

---

인대파열을 겪고계시지는 않으신가요?인대는 유연성과 인장강도가 뛰어나 잘 늘어나지 않아 약해져있으면 고정시켜주기가 쉽지 않습니다.약하게 고정해주시고 무릎을 꿇으시거나 걸어다니는 정도가 높다면 이완시켜주셔야할 필요가 있습니다.인댄 상태에서 체중작업시에는 무리하지 않게 해주셔야 합니다.

---

### PPO

---

원인이 정확하게 일치하지는 않겠지만요. 아킬레스 건염이 있으신가요? 아킬레스 건염은 스포츠 재활 할때 치료해야 할 것으로 생각합니다. 기억하시기 바랍니다. 감사합니다.

---

## ABSTRACT

# Application of Reinforcement Learning to Response Generation Model Using Domain Data

ByeongCheol MIN

Department of Big Data

Sungkyunkwan University

Since the release of ChatGPT, many people have been experiencing the performance of large-scale language models and sharing various ways to utilize them through social media. Most of them are In-Context Learning. But utilizing In-Context Learning is insufficient for use for a specific purpose. Therefore, a model trained with other types of data is needed. There are practical challenges to train model including the need for specialized personnel, infrastructure, time, and costs. OpenAI offers an API to train GPT models, but there is a security problem because it requires internal company data to be leaked to the outside.

This study proposes an approach to building a generative model applying the ChatGPT training method, InstructGPT, to achieve specific business objectives and evaluates their performance. The scope of the research is limited to the task of generating responses to patients' queries using medical counseling data. In addition, instead of the HumanFeedback method which ranks the sentences

that people prefer among the various responses generated by the fine-tuning model, the ranking was based on a score indicating the degree of expert activity.

As the result of the study, the responses generated by the fine-tuning model and the reinforcement learning model were each measured for similarity to the expert's response, and it was confirmed that the reinforcement learning model improved performance.

Although there was significant reward variance during the reinforcement learning process, it was confirmed that learning progressed based on various metrics. Furthermore, when inspecting responses, it was observed that fine-tuning model often generated repetitive words or left sentences unfinished, whereas reinforcement learning model showed better performance in these aspects.

**Keyword: Generative Model, LLM, Reinforcement Learning, Q&A, Natural Language Preprocessing**