

WERODA



챗봇 # 심리상담 # KoGPT2 # Elasticsearch

NLP-10
(핫식스)

CONTENT^o

Intro. 팀 소개 / 프로젝트 소개	01
Model. 데이터셋 / 모델 / 최종 적용 모델	02
Product Serving. 서비스 아키텍처	03
Result. 시연 영상 / 결과 및 고찰	04
Appendix. 도전적인 실험 / Q&A	05

Intro



김남현_T3021

- Hate Speech Filtering
- Data Collection



민원식_T3079

- PM
- Validation Server
- Data Collection



전태양_T3194

- EDA
- Data Collection



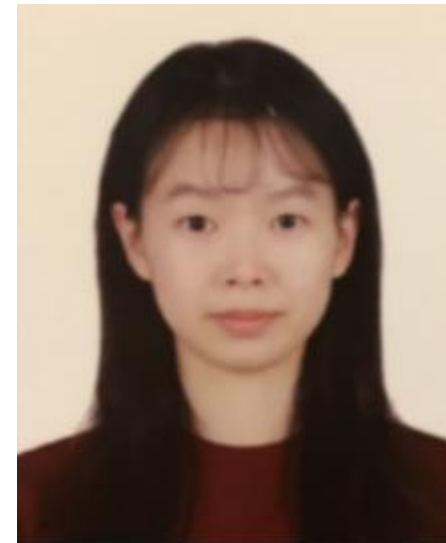
정기원_T3195

- Elastic Search
- Text Style Transfer
- Data Collection



주정호_T3211

- Model Research
- UI
- Server Deploy

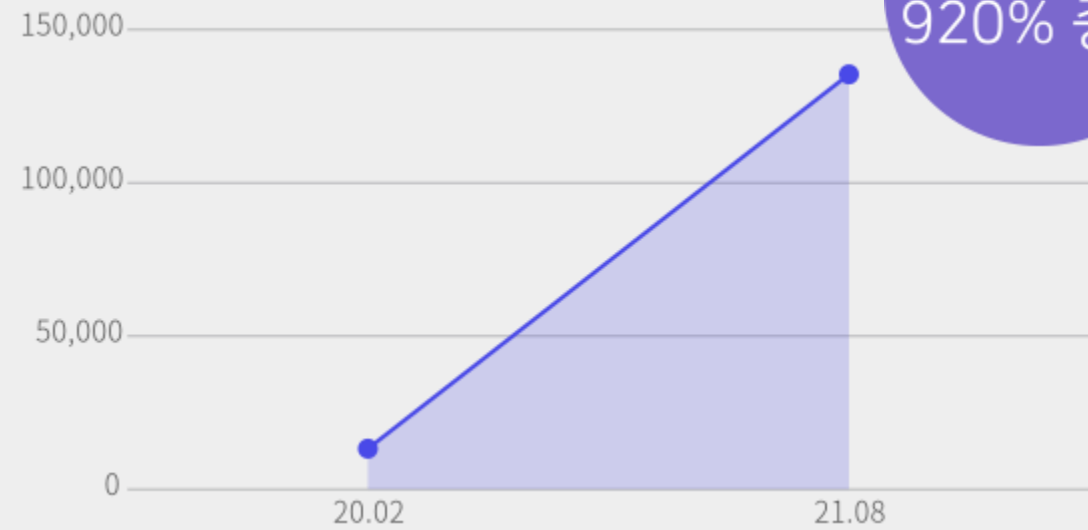


최지민_T3223

- DialogBERT
- Data Collection

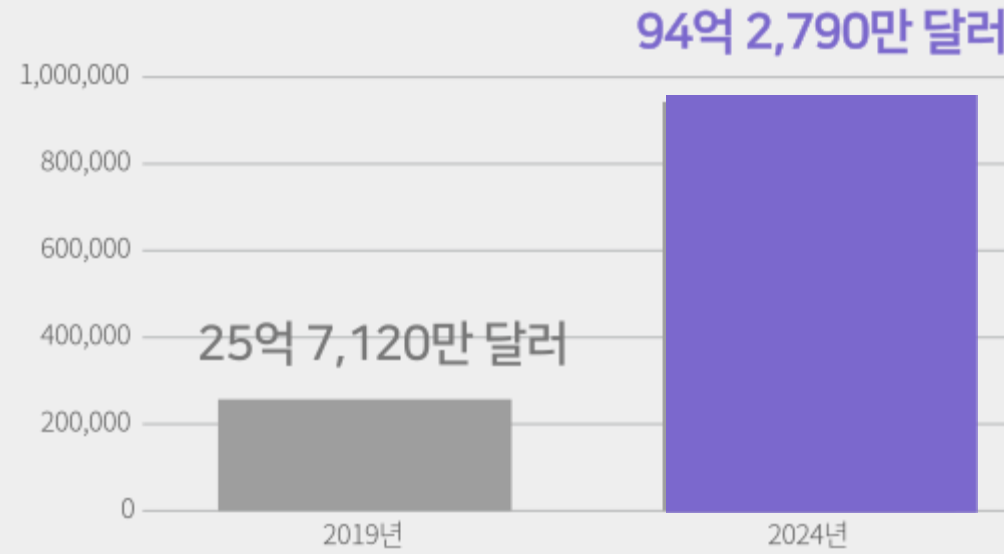
주제 선정 배경

[개인 심리상담 횟수] (단위 : 건)



(출처 : 코로나 19 통합 심리 지원단)

18개월 간
920% 증가



(출처 : 연구개발특구 진흥공단)

20년 2월에 13,265건에서 21년 8월에 135,407건으로 개인 심리상담 횟수가 무려 920% 증가했다.

이는 심리상담 서비스에 활용되고 있는 챗봇 서비스에 대한 수요가 앞으로도 커진다는 점을 시사한다.

2019년 기준으로 챗봇 시장 가치의 연평균 성장률은 29.7%로 2024년에는 약 11조원에 이를 것으로 전망된다.

기존 서비스의 한계

[기존 심리상담 챗봇]

개발사	이름
과학기술정보통신부 & 한국연구재단	마이멘탈 포켓
서울시 COVID19 심리지원단	누구나 챗봇
아모터스	마인드카페

"우울증 치료? 복장 터져"... 디지털치료제 시범서비스 직접 써보니

과기정통부, '마이멘탈포켓' 1일부터 시범 서비스
 인공지능 챗봇 '상담' 코너, 영동한 답변 속출
 다른 서비스는 아직 준비중

['마이멘탈 포켓'의 낮은 사용자 만족도]



['마이멘탈 포켓'의 단발성 답변]



['누구나 챗봇'의 트리형 구조]

심리 상담원과 직접 대화할 수 있는 챗봇은 주로 유료 서비스이고, 무료 심리상담 챗봇은 단발성 또는 트리형 구조인 경우가 많다.

개발 목표 및 기대 효과

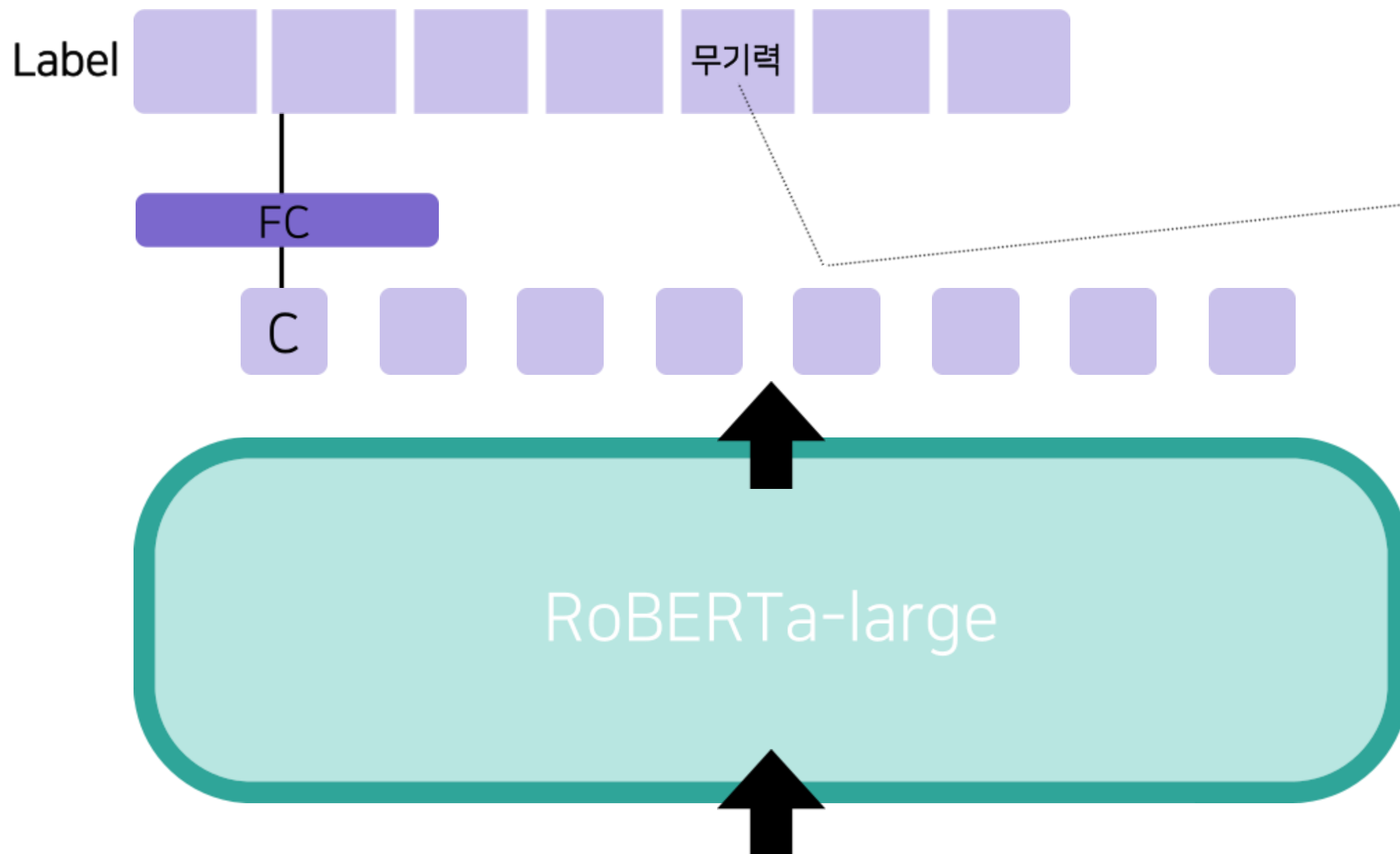
<p>트리형 구조 NO</p>  <p>트리형 구조는 사용하지 않습니다.</p>	<p>단발성 대답 NO</p>  <p>단발성 대답은 사용하지 않습니다.</p>	<p>문맥에 맞는 대화</p>  <p>멀티턴 대화 기능 일상 대화형 챗봇</p>
<p>정신 건강 케어</p>  <p>사용자의 상황 고려 공감과 위로의 채팅</p>	<p>인건비 절감</p>  <p>심리상담원 NO 딤러닝 YES</p>	<p>부가 서비스</p>  <p>사용자의 성향 분석 음악 및 상품 추천</p>

Model

데이터셋

이름	특징	비고
[KETI] 웰니스 대화 스크립트 데이터셋	<p>데이터셋1: 정신 건강 상담 주제의 359개 대화 의도에 대한 5,232개의 사용자 발화 및 1,023개의 챗봇 발화</p> <p>데이터셋2: 사용자-챗봇의 대화쌍을 turn에 맞추어 정신 건강 상담 주제의 19개 대화 의도에 대한 발화로 재구성 챗봇의 답변에 따른 사용자 발화 중 '응', '아니'와 같은 답변을 <긍정답변> <부정답변>의 label로 전환</p>	KoGPT2 학습, Retrieval
[songys/Chatbot_data] Chit-Chat 데이터셋	<p>다음 카페 "사랑보다 아름다운 실연"에서 자주 나오는 이야기들을 참고하여 제작 챗봇 트레이닝용 문답 페어 11,876개 일상 0, 이별(부정) 1, 사랑(긍정) 2로 라벨링</p>	Retrieval
[AI-HUB] 감성 대화 말뭉치	<p>60가지 다양한 감정의 코퍼스 27만 문장 우울증 관련 및 대화 응답 시나리오 포함 감성 대화 엔진 또는 챗봇을 개발하려는 목적에 맞추어 제작 문장에서 의미와 의도를 추출하는 확률이 기존 통계 모델링 기법(CRF+ 등)에 비해 월등히 높은 성능을 보여줌</p>	Retrieval

모델 - RoBERTa

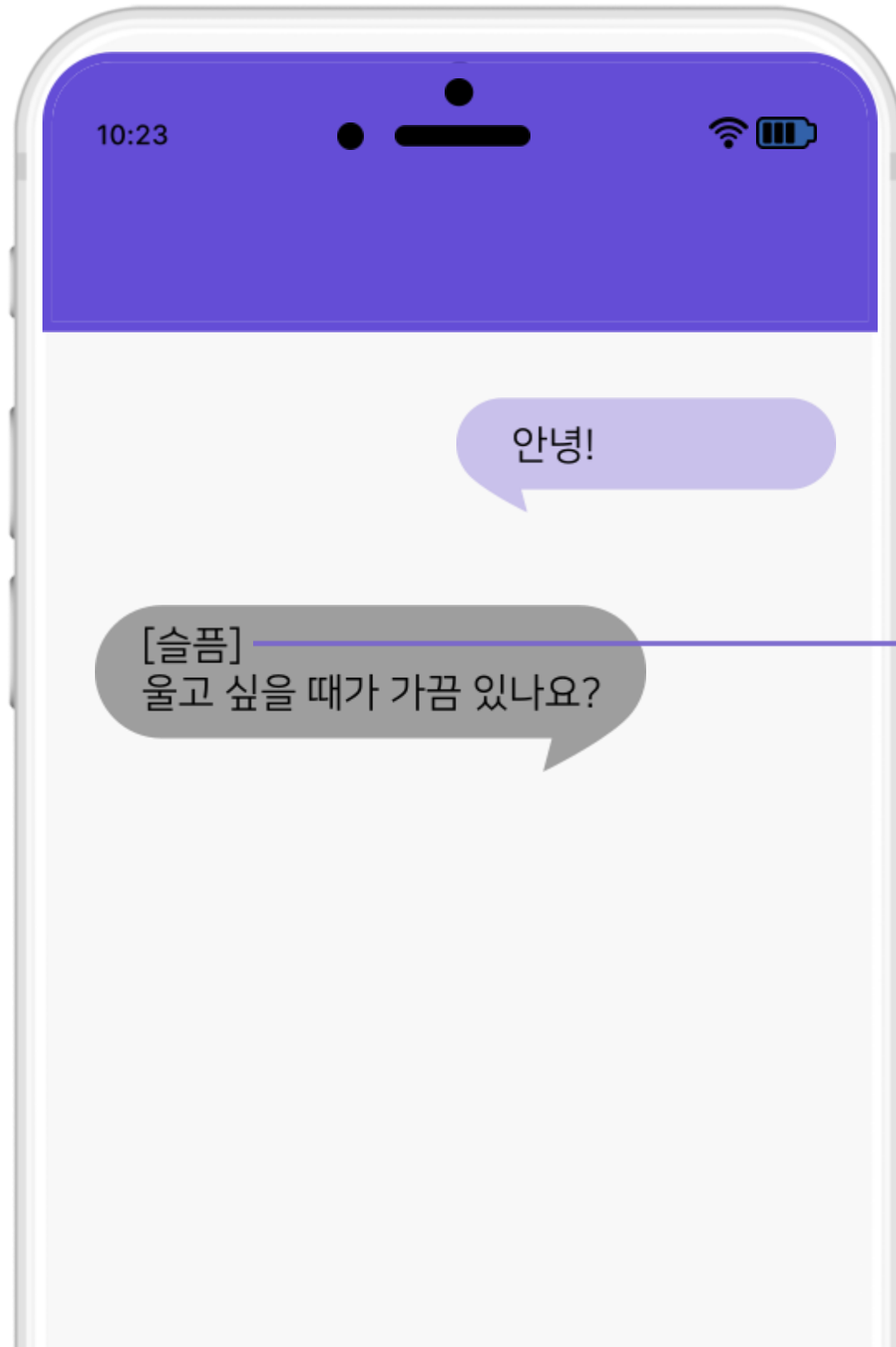


[CLS] 하루종일 공부했더니 몸이 피곤하네 [SEP]

Wellness Dataset

심리상담 주제	답 변
...	...
무기력	잠시 휴식을 취하는 것이 좋겠어요. ✔ 따뜻한 물에 목욕하는건 어때요? ...
...	...

모델 - RoBERTa

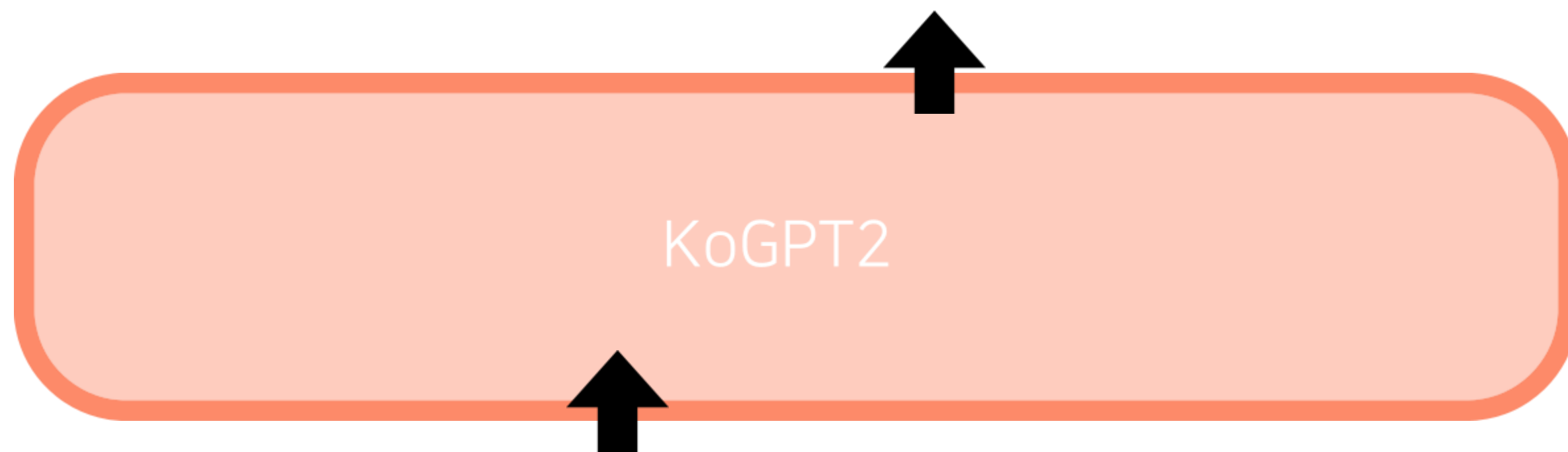


일상 대화를 심리상담 주제로 분류

→ 일상 대화 불가능

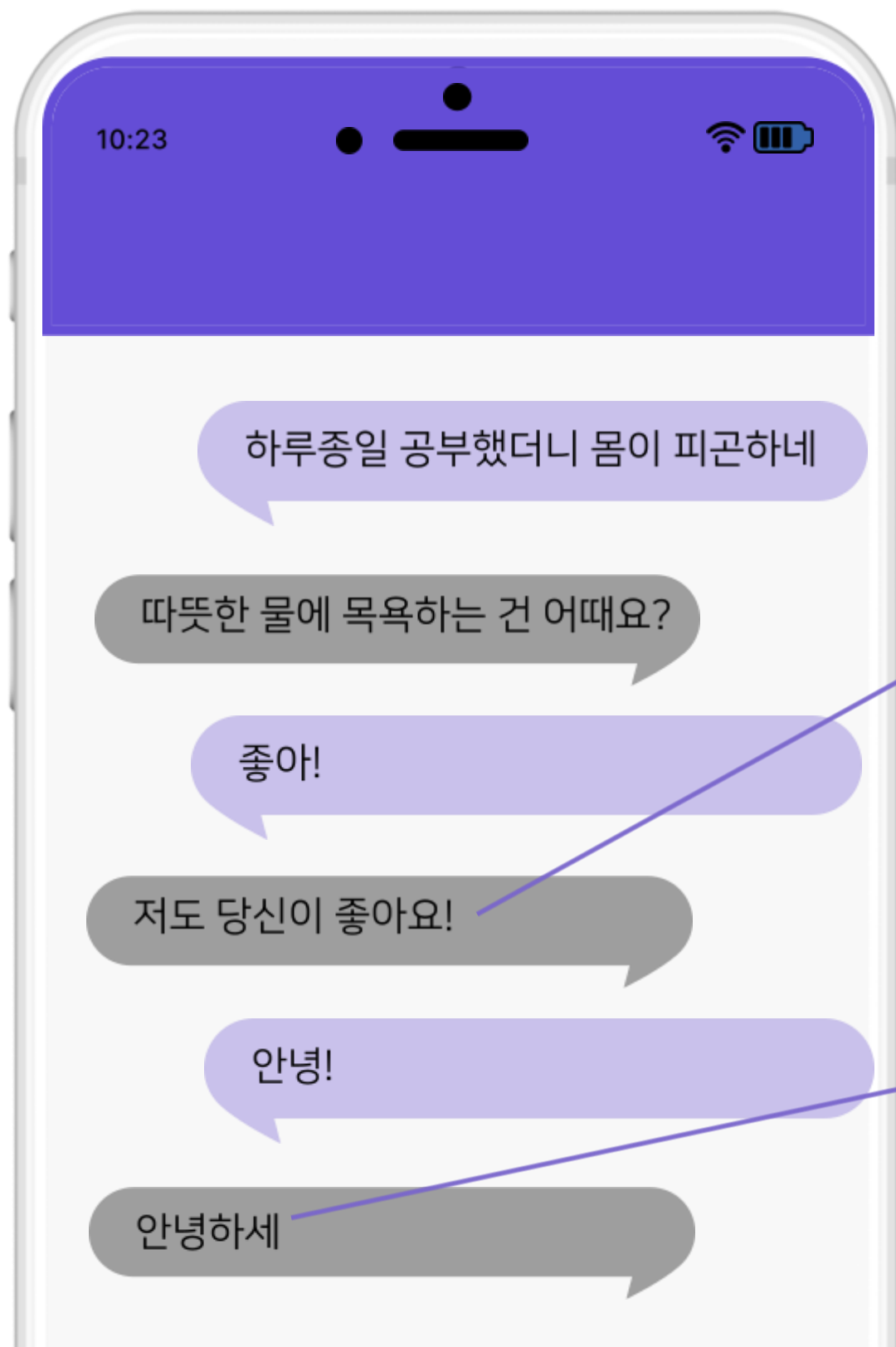
모델 - KoGPT2 (Single-Turn)

따뜻한 물에 목욕하는건 어때요? </s>



<s> 하루종일 공부했더니 몸이 피곤하네 </s> <s>

모델 - KoGPT2 (Single-Turn)



챗봇이 이전 대화를 기억하지 못함

→ 연속된 대화 불가능

답변의 완성도가 떨어지는 생성 모델의 특성

→ 챗봇의 성능 하락

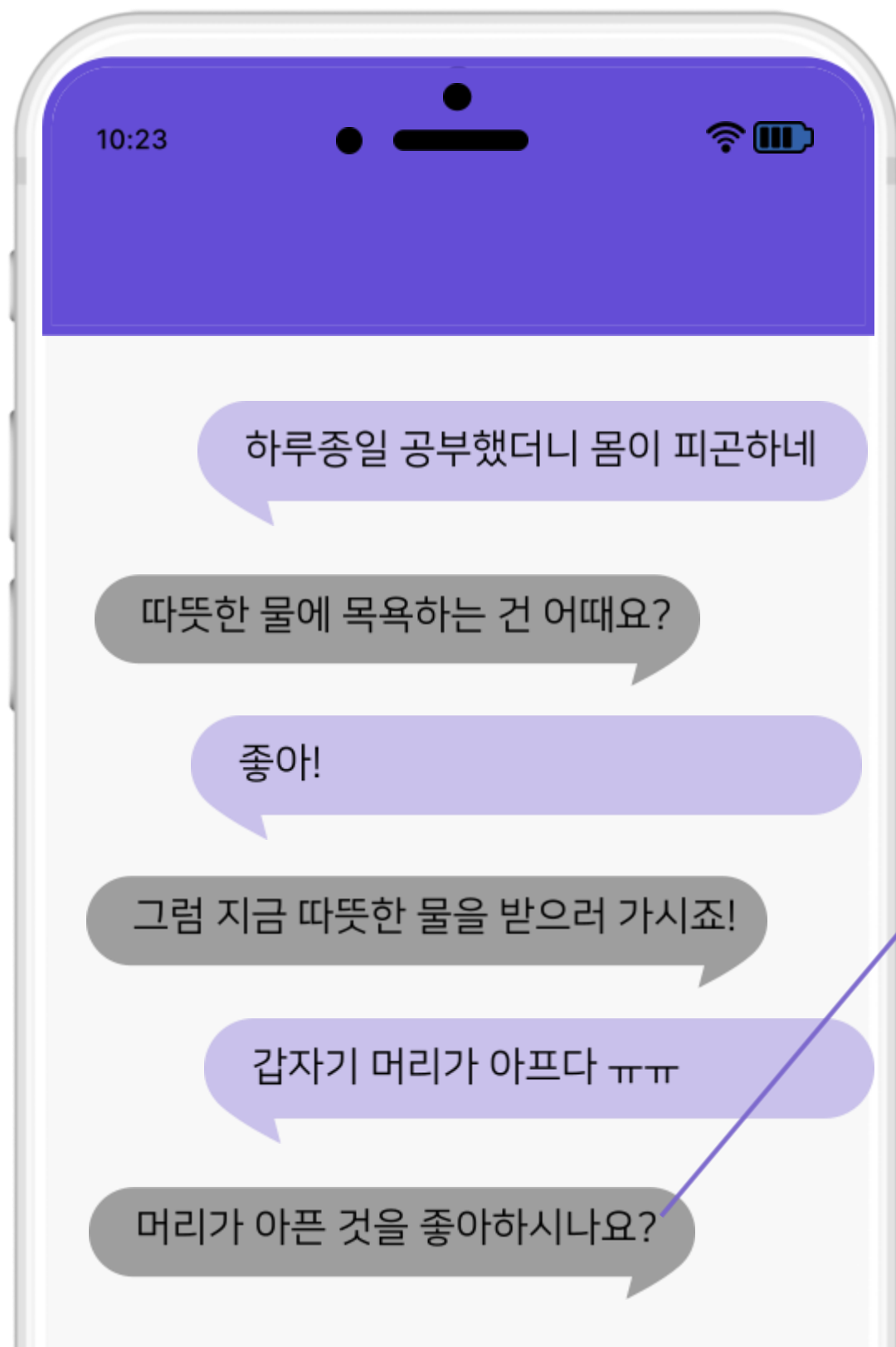
모델 - KoGPT2 (Multi-Turn)

그럼 지금 따뜻한 물을 받으러 가시죠! </s>



<s> 하루종일 공부했더니 몸이 피곤하네 </s> <s> 따뜻한 물에 목욕하는건 어때요? </s> 좋아! </s> <s>

모델 - KoGPT2 (Multi-Turn)



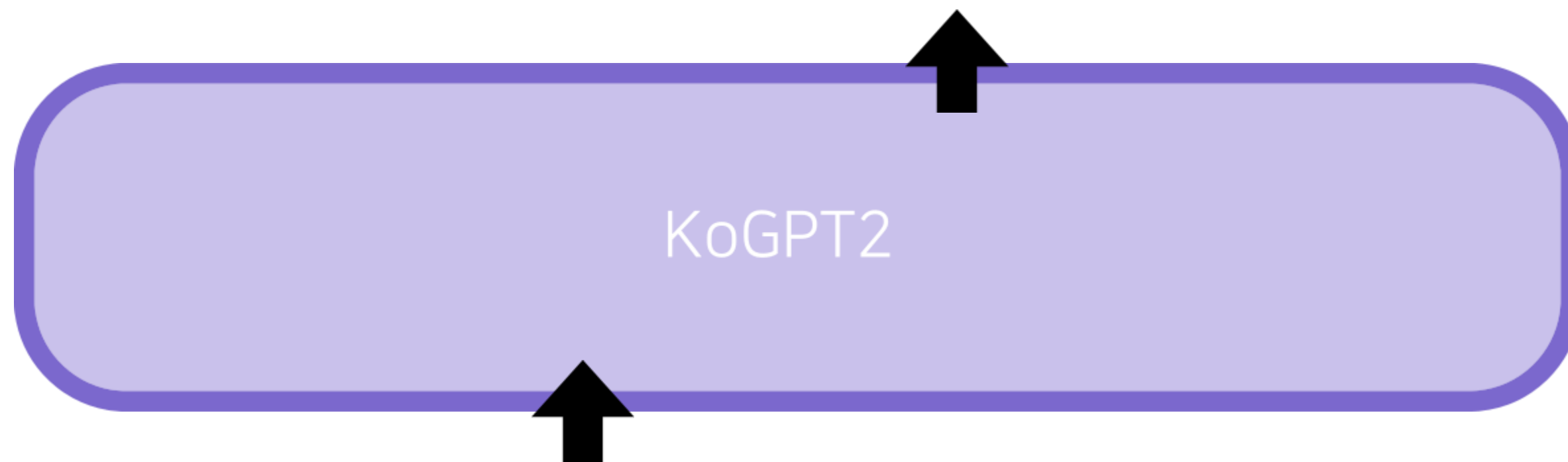
이전 대화와 상관없는 메시지를 입력해도, 이전 대화를 참고하여 답변 생성

→ 새로운 맥락의 대화는 대응 불가능

모델 - KoGPT2 (Multi Downstream-Task)

classification + dialog

무기력 <sys> 따뜻한 물에 목욕하는 건 어때요? </s>



<usr> 하루종일 공부했더니 몸이 피곤하네 <sent>

모델 - KoGPT2 (Multi Downstream-Task)

user	system	sentiment
너무 불안한거 있지.	불안감을 느끼고 계시는군요. 00님이 불안감으로 많이 힘드신 건 아닌지 걱정이 돼요.	불안
완전 힘들고 짜증나요.	평소 즐기던 일이 더 이상 즐겁지 않거나 흥미가 사라졌나요?	무기력
다 때려치고 싶어.	마음이 잠깐 지쳤나 봐요. 아무 의미 없는 일은 없을 거예요.	의욕 상실
어 맞아	그러셨군요. 많이 당황스러우셨을 것 같아요. 이런 상황이 오래 지속될까 걱정돼요.	긍정답변
그건 아닌데	그렇군요. 불안을 줄이는 방법으로 꾸준한 운동이나 명상, 규칙적인 수면이 도움이 될 수 있어요.	부정답변

[KETI] Wellness Dataset

user	system	sentiment
12시 땡!	하루가 또 가네요.	일상
안녕	안녕하세요!	일상
3박4일 놀러가고 싶다.	여행은 언제나 좋죠!	일상
반가워!	저도 반갑습니다!	일상
넌 누구야?	저는 '핫식스' 팀에서 개발한 위로봇입니다.	일상

[songys/Chatbot_data] Chit-Chat Dataset

Auto-Regressive

무기력 <sys> 따뜻한 물에 목욕하는 건 어때요? </s>

KoGPT2

<usr> 하루종일 공부했더니 몸이 피곤하네 <sent>

일상 <sys> 도움이 되어서 뿌듯하네요! </s>
일상 <sys> 다른 방법을 생각해볼게요. </s>

Multi-Turn

KoGPT2

<usr> 하루종일 공부했더니 몸이 피곤하네 응, 좋은 생각이다! <sent>

<usr> 하루종일 공부했더니 몸이 피곤하네 아니, 별로인데? <sent>

↑
이전 발화를 추가

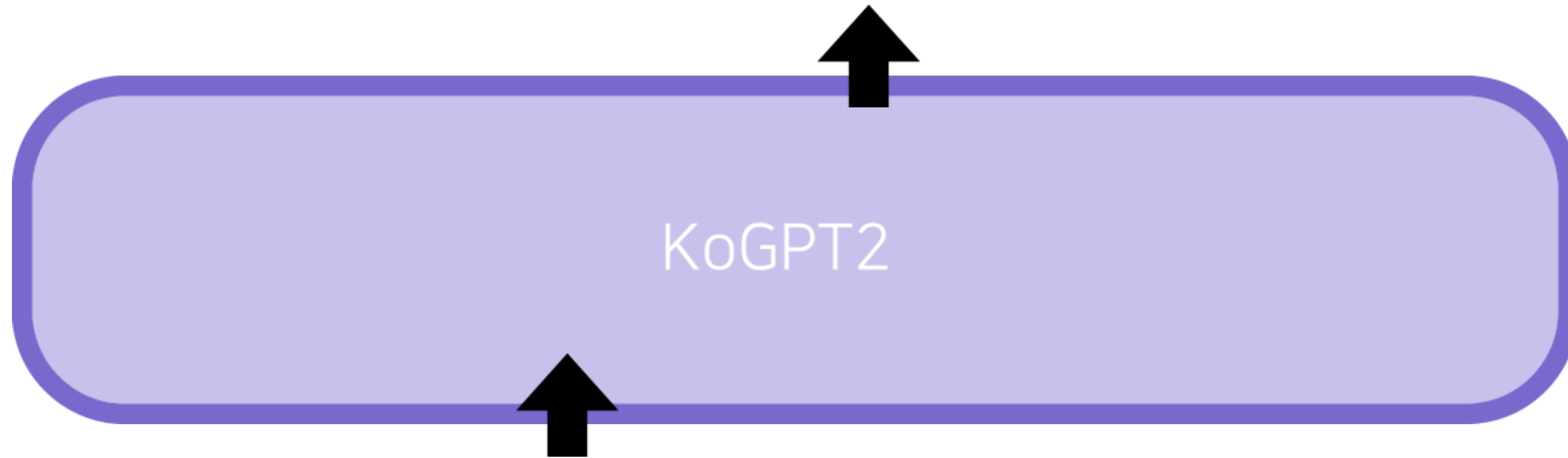
↑
긍정답변
부정답변

KoGPT2

<usr> 응, 좋은 생각이다! <sent>

<usr> 아니, 별로인데? <sent>

Beam-Search { 초조함 <sys> 많이 초조하신가요? </s>
무기력 <sys> 따뜻한 물에 목욕하는 건 어때요? </s> ✓
무기력 <sys> 오늘은 휴식을 취하는 것이 어떤가요? </s>
외로움 <sys> 마음이 많이 지쳤나봐요 </s>
무기력 <sys> 온 몸에 힘이 다 빠진 느낌인가요? </s>



<usr> 하루종일 공부했더니 몸이 피곤하네 <sent>

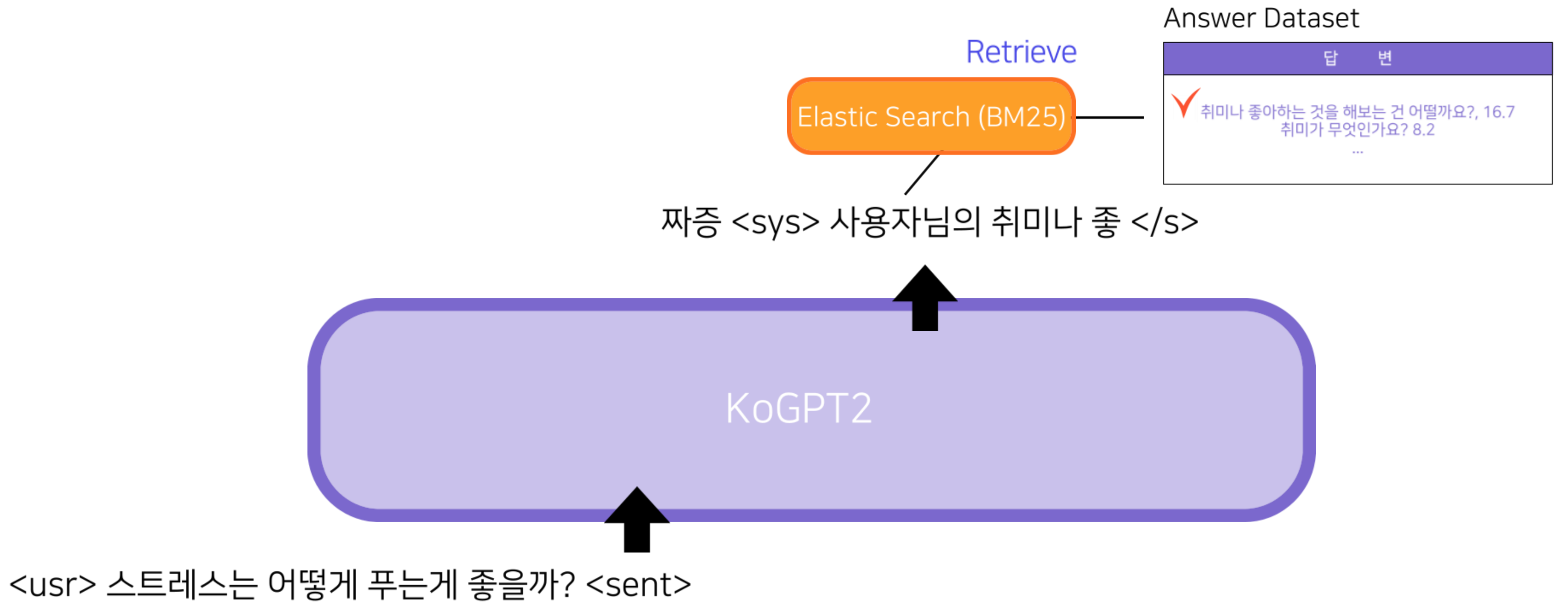
답변의 완성도가 떨어지는 생성 모델의 특성

→ 챗봇의 성능 하락

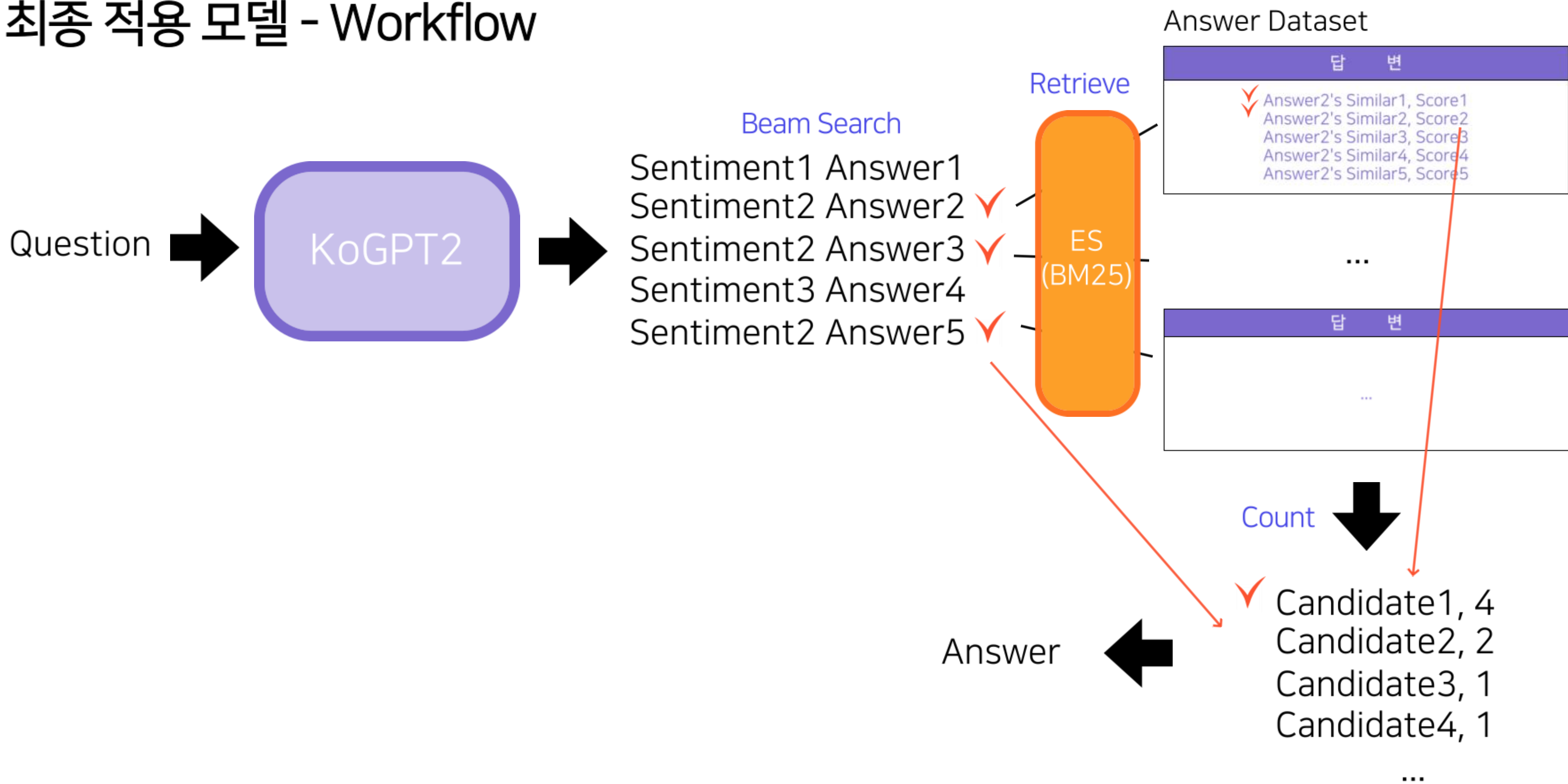
짜증 <sys> 사용자님의 취미나 좋 </s>

KoGPT2

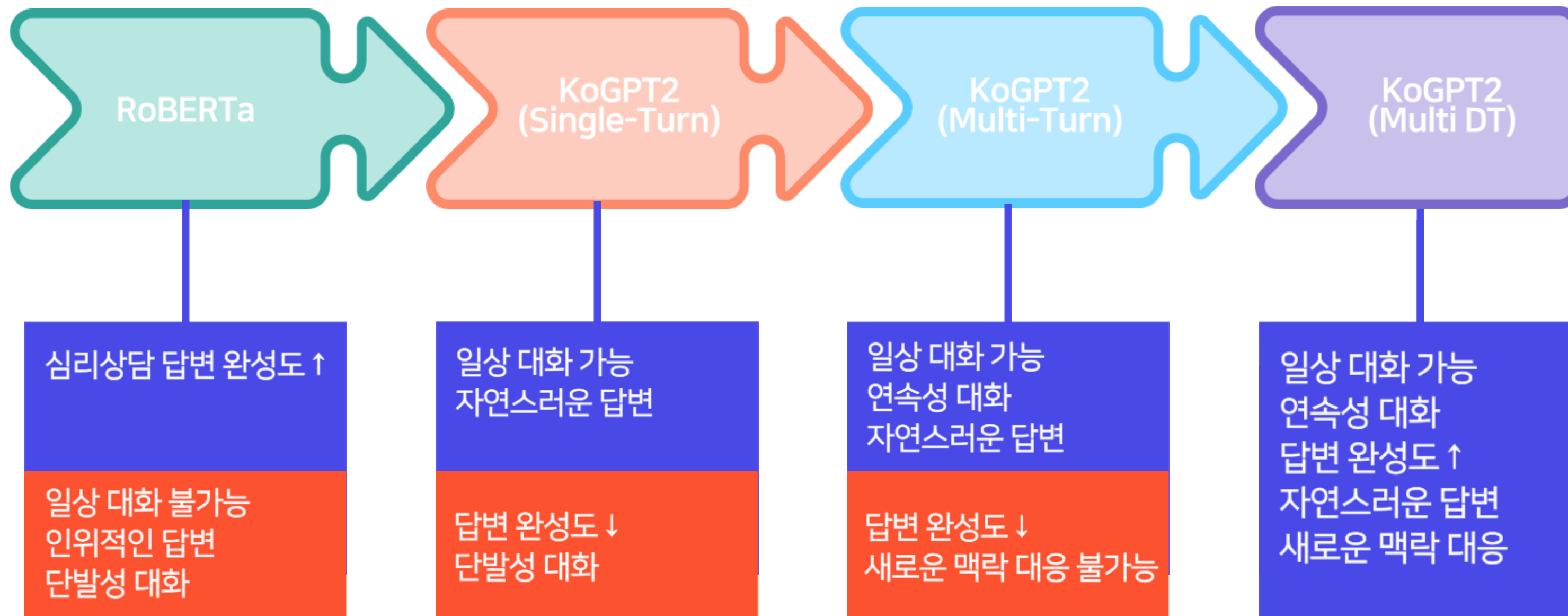
<usr> 스트레스는 어떻게 푸는게 좋을까? <sent>



최종 적용 모델 - Workflow



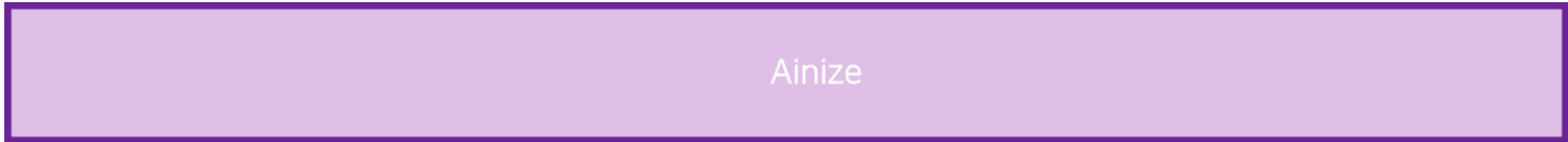
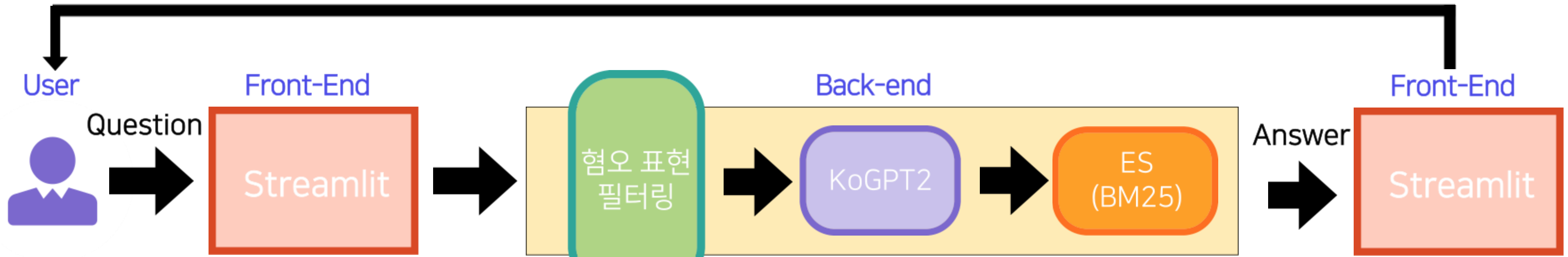
최종 적용 모델 - Timeline



Product Serving

서비스 아키텍처

1 second or less



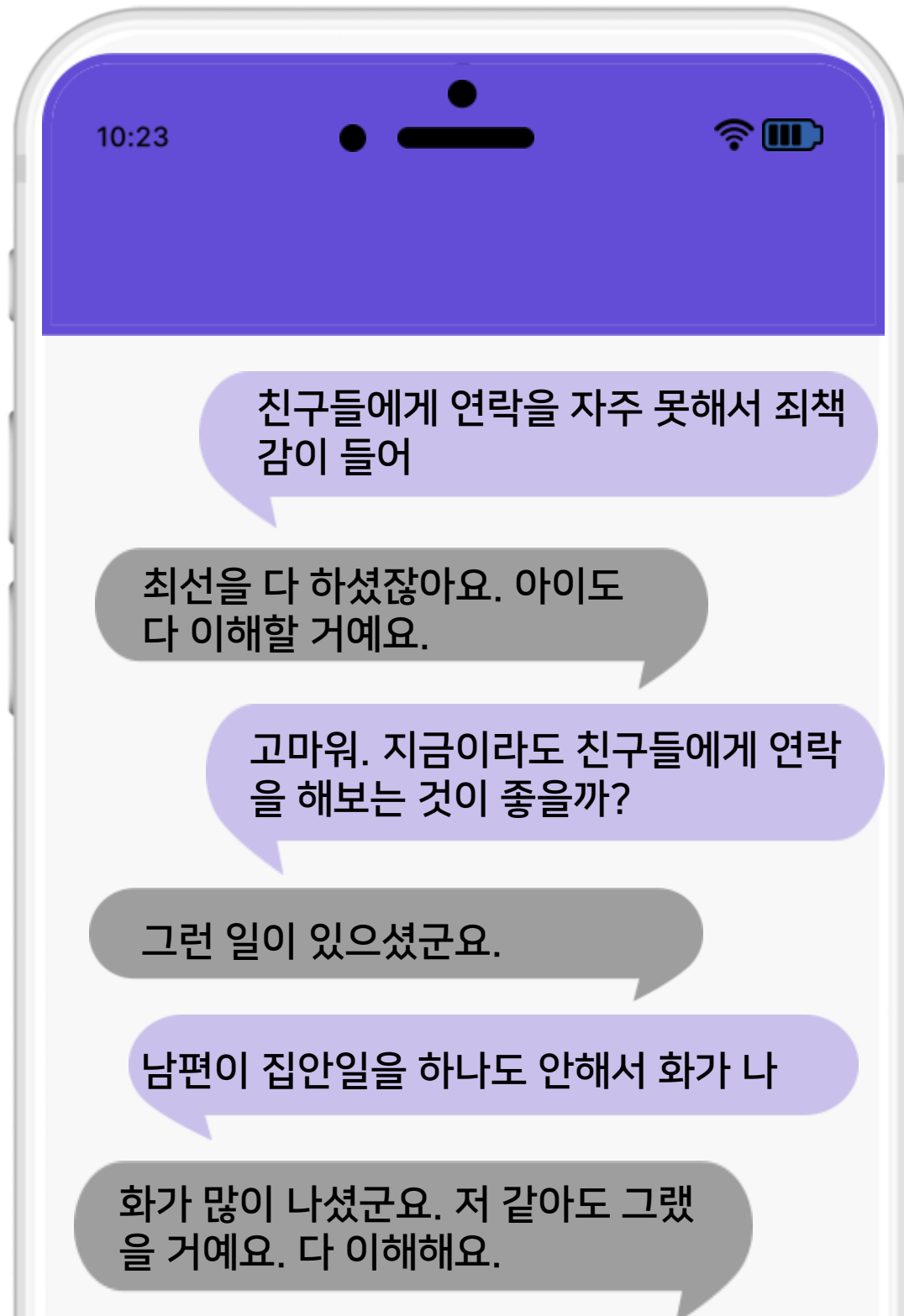
Server

Result

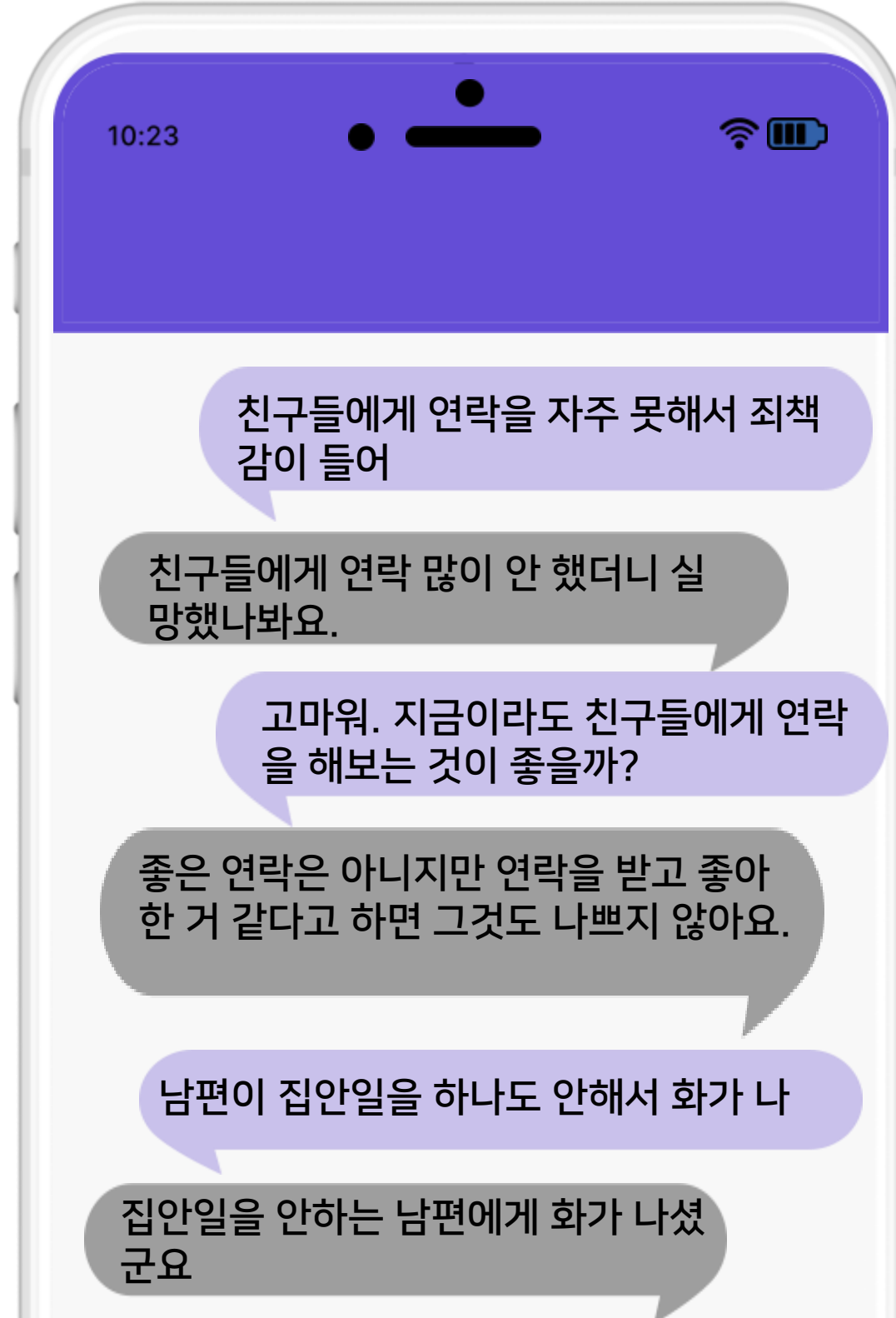


Demo site: [Streamlit Cloud](#) [Ainize](#)

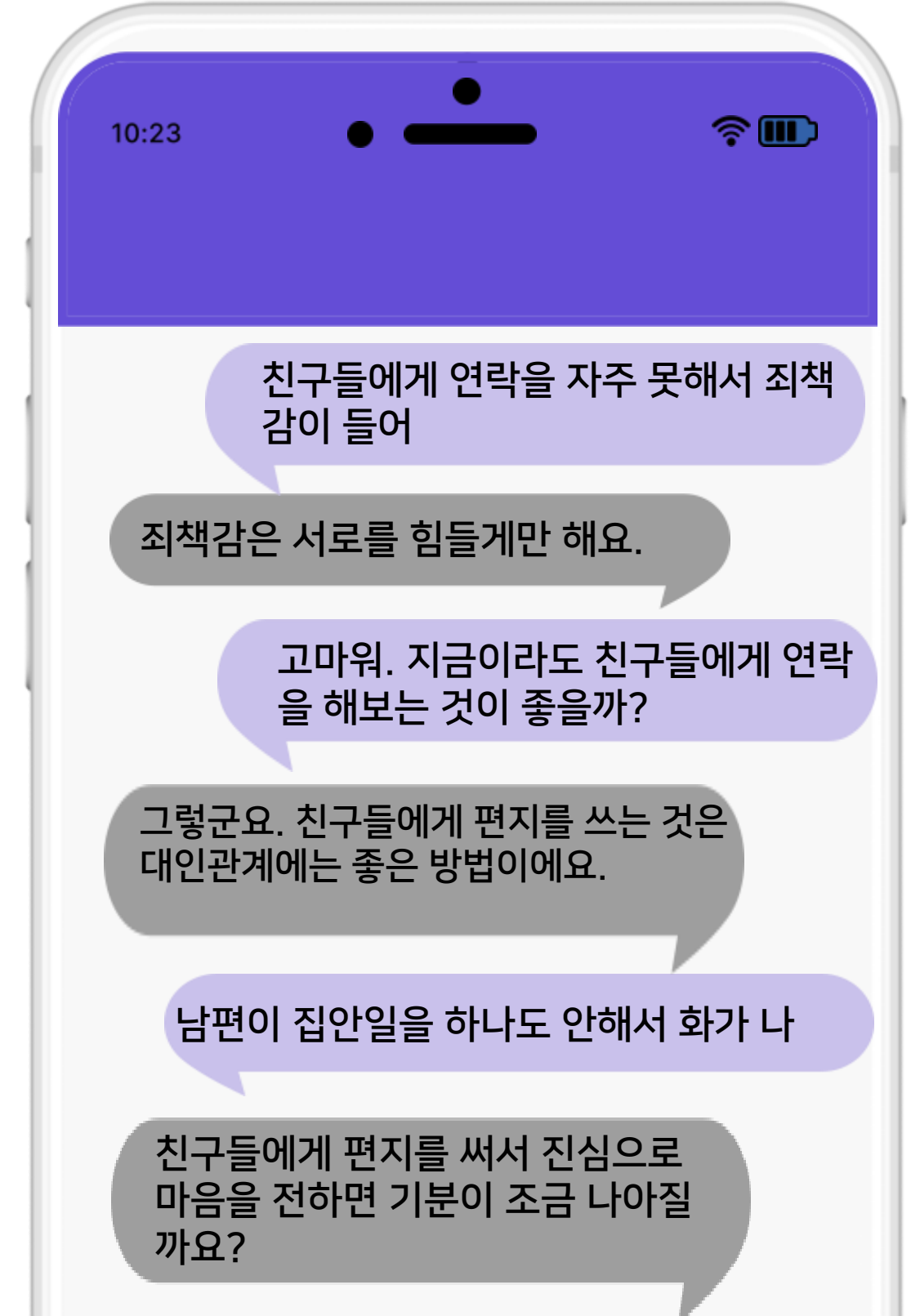
모델 - RoBERTa



모델 - KoGPT2 (Single-Turn)



모델 - KoGPT2 (Multi-Turn)



- SSA(Sensibleness & Specificity Average) 지표란?
 - 2020년 구글이 챗봇 Meena를 발표하면서 도입한 대화 만족도 평가 지표
 - Sensibleness: 답변의 맥락과 논리성 평가
 - Specificity: 답변의 구체성과 사람과의 유사성 평가
 - e.g. Q) 너 중국 음식 좋아해?
 - A1) 오늘 메뉴가 뭔가요? (Sensibleness: 0, Specificity: 0)
 - A2) 응, 좋아해 (Sensibleness: 1, Specificity: 0)
 - A3) 응, 난 그 중에 '짜장면'이 제일 좋아. (Sensibleness: 1, Specificity: 1)

다수의 대화 만족도 평가자들이 평가한 지표의 평균이 바로 SSA 수치이다.

결과

평가자 간 일치도 평가

- 최종 모델의 성능 평가자 간 일치도 평가란?

- 구글 Meena 발표 논문의 작업자 간 일치도 평가 방법을 참조

- Krippendorff's alpha: 0~1 사이의 값 (0: 불일치, 1: 일치)을 가짐

- Agreement: SSA 계산을 위해 사용

Metric	Agreement (%)	Krippendorff's alpha
Sensibleness	66.6	0.74
Specificity	59.3	0.67

[Our Team Worker Agreement]

Metric	Agreement (%)	Krippendorff's alpha
Sensibleness	76 ± 3	0.42 ± 0.03
Specificity	66 ± 2	0.30 ± 0.05

[Meena's Crowd Worker Agreement]

작업자 간 일치도 평가 수치(Krippendorff's alpha)가 Meena 논문보다 높은 값을 달성하였다.

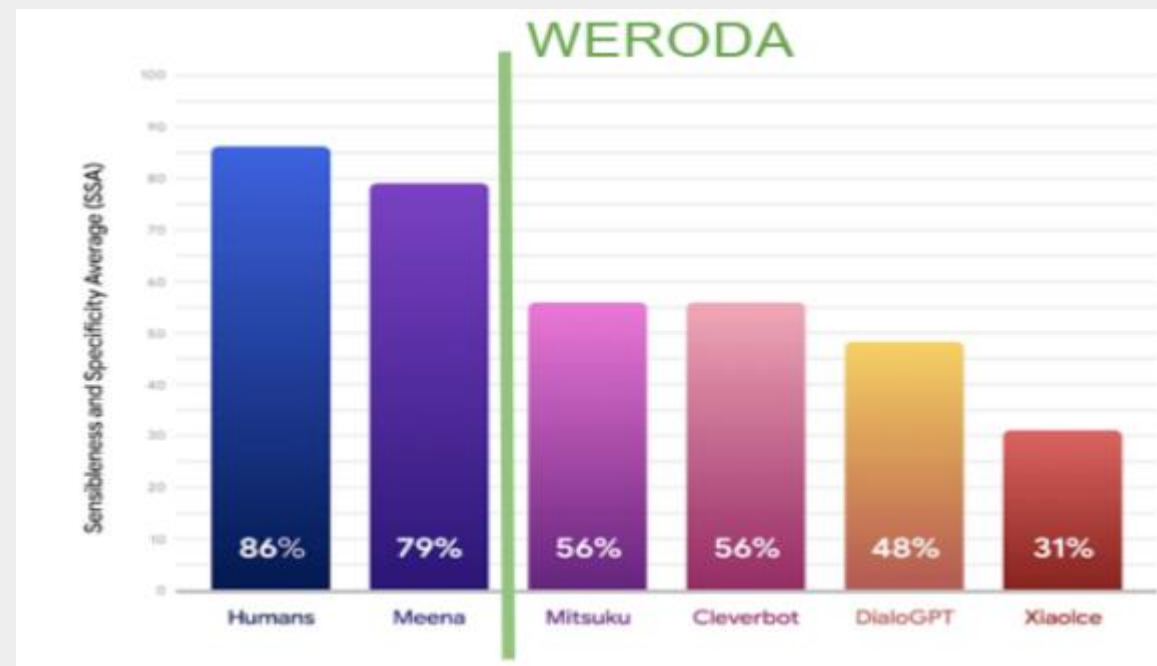
결과

모델 성능 측정 및 비교

- 모델 성능 측정 [BentoML을 이용해 검증 서버를 구현하여 수행 / 팀원 6명이 각각 50개의 대화, 총 300개의 대화쌍으로 평가]

Model	Sensibleness	Specificity	SSA	비 고
RoBERTa	.	.	.	평가 불가
KoGPT2 (Single-Turn)	42.13	38.49	40.31	.
KoGPT2 (Multi-Turn)	39.39	34.57	36.98	.
KoGPT2 (Mutl Downstream-Task)	66.60	59.26	62.93	최종 모델

- 타 모델과 비교



고찰

- 최근에 런칭한 루다와 같은 상용 챗봇은 다양한 주제에 대해 대화를 나눌 수 있어 성능 평가 비교의 대상으로 삼기 어려운 점이 있으나, 심리 상담 주제로 한정한다면 정부의 챗봇에 비해 사용자의 만족도가 향상되었음
- 다만 자연스러운 대화를 위해 더 많은 DataSet을 확보하는 부분과 각 사용자에 적합한 챗봇의 다양한 페르소나를 반영하는 것이 당면 과제

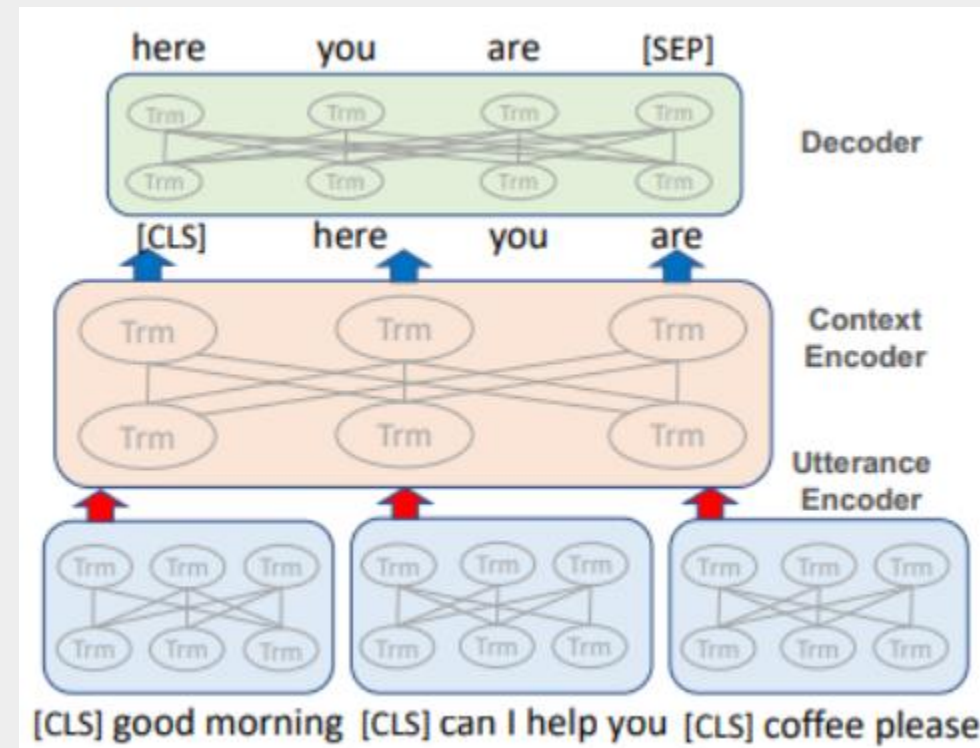
후속 연구 및 개발 목표

- 일상 대화 및 심리 상담과 관련된 더 많은 학습 DataSet 확보
- 사용자와 자연스러운 대화를 위해 챗봇에 페르소나 도입
- Retrieval 후보를 의미에 기반하여 선정하도록 하여 답변의 정확도 향상(현재 BM25 문서유사도에 의해 후보 선정)
- 챗봇 성능 평가 시 정확성을 높이기 위한 평가 가이드라인 구체화
- 사용자 취향, 취미 등을 바탕으로 맞춤형 서비스 제공에 관한 기능 구현 및 해당 서비스로 발전

Appendix

DialogBERT

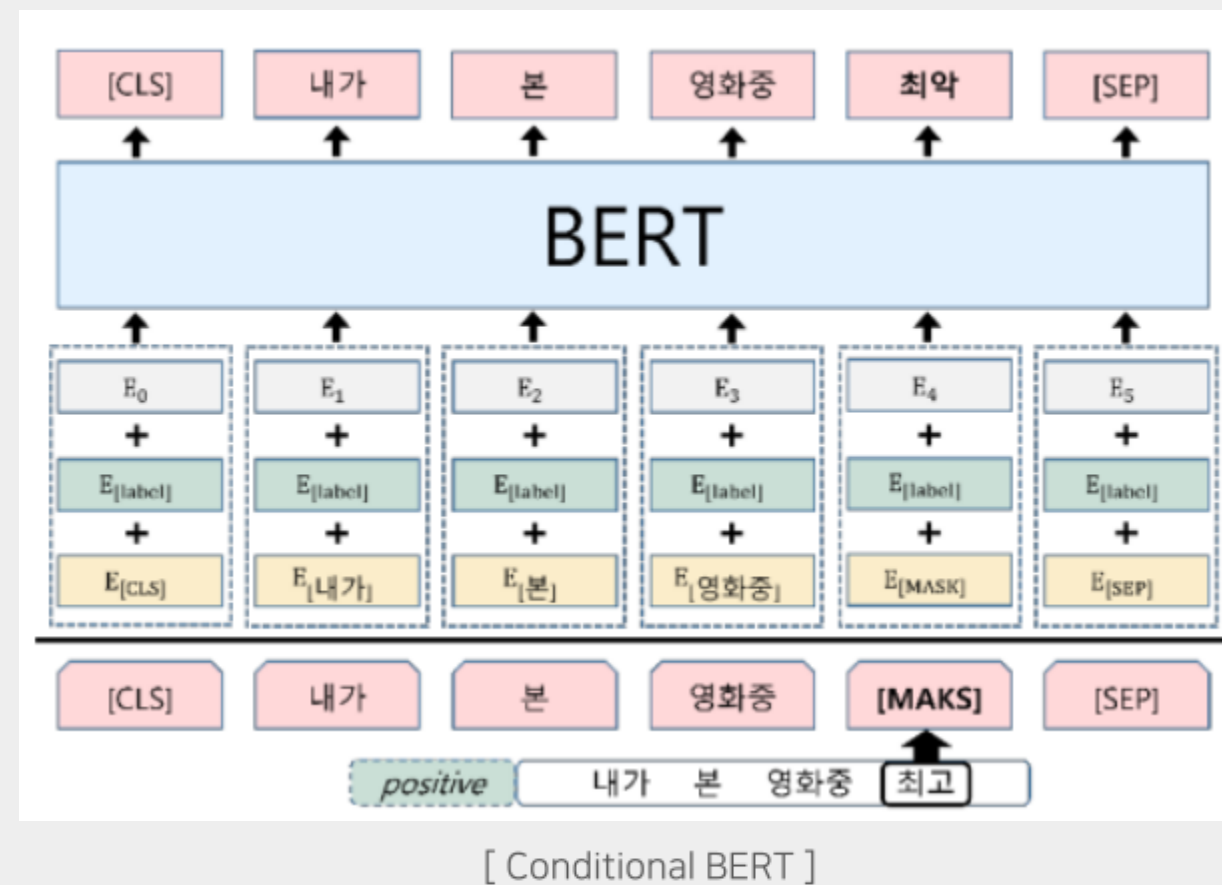
- 연속성 대화를 위해 시도한 다른 방법
- Model Structure
 - Utterance Encoder: 각각의 발화를 인코딩
 - Context Encoder: 인코딩된 발화에 대한 self attention을 통해 맥락을 고려한 representation 생성
 - Decoder: 맥락 representation을 이용해 다음 발화 생성
- Training Objectives
 - Next Utterance Generation: Decoder가 생성한 단어와 맥락 representation을 조건부로 다음 토큰 예측
 - Masked Utterance Regression: 마스킹된 발화에 대한 벡터를 reconstruct
 - Distributed Utterance Order Ranking: 무작위로 섞인 발화문을 원래 순서로 맞추는 작업



Text Style Transfer (연령대 별)

- 챗봇 사용자의 연령대는 다양하지만, 답변의 말투가 일관적이라 아쉽다.
→ 답변의 말투를 변환하여 친근감 있고 재밌는 챗봇을 만들자!

- e.g. Q) 다음 시험도 노력이 결실을 보길 바라게요.
A1) 다음 시험도 노후 결실을 보길 바라게요. (노년)
A2) 다음 시험도 노력 결실을 보기만 바라게요. (청소년)



Q&A