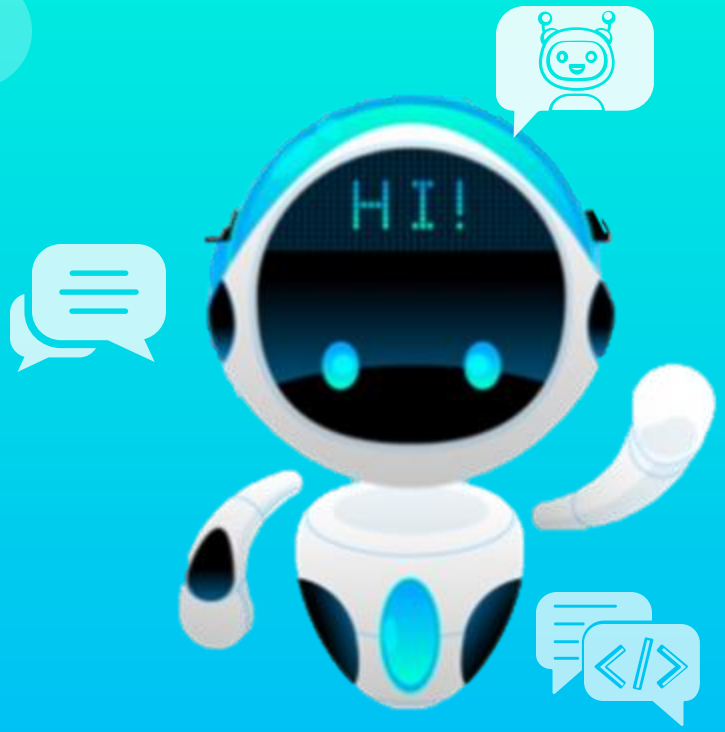
ERU-JIMYOUNG



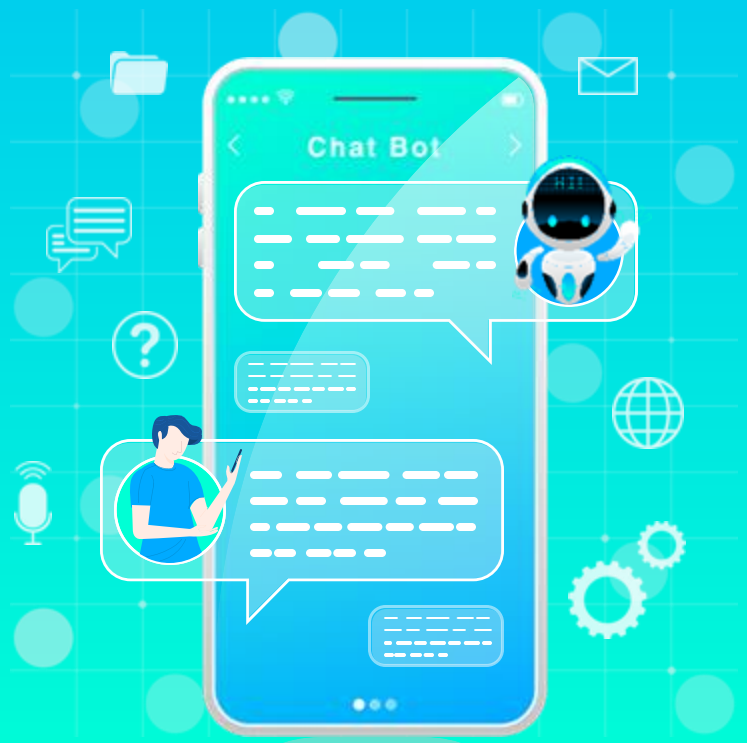
MBTI
Chat Bot

KDT 이루지명

백지명 강진영 고예성 박현식 안영준 조세연

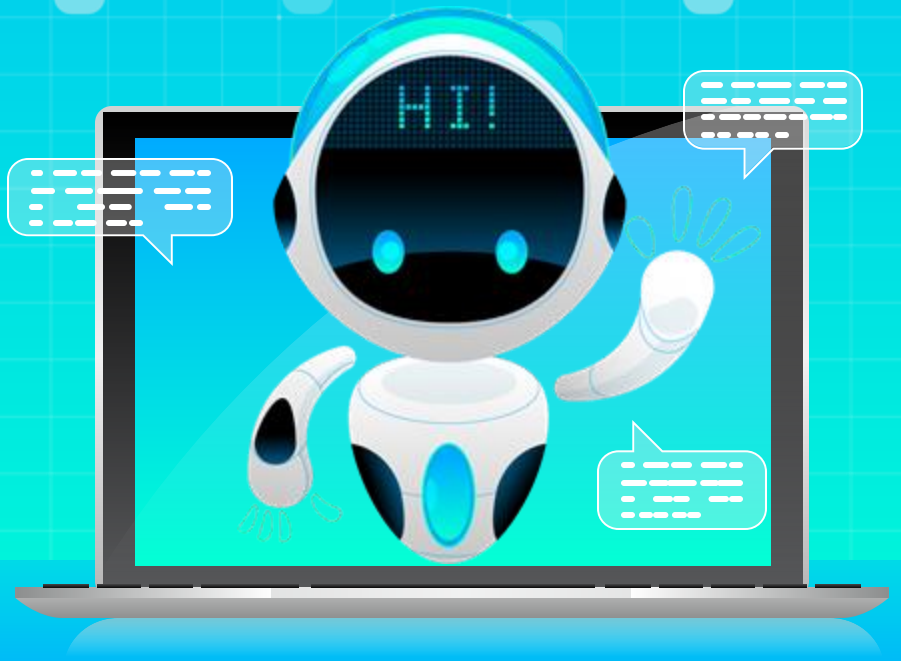
MBTI TALK

CONTENTS



01

서비스 제작 배경



1. 서비스 제작 배경

MBTI

01

MBTI의 관심도 상승

- MBTI는 최근 가장 인기 있는 개인 성격 검사 도구
- MBTI의 지표 중 특히 T/F의 차이가 가장 사람들의 흥미를 유발

02

사용자 맞춤 서비스

- 개인의 대화 성향을 파악하여 더 나은 대화가 이루어짐
- 각기 다른 MBTI나 소통 방식으로부터 벗어나 자신의 대화 방식에 맞는 소통 가능

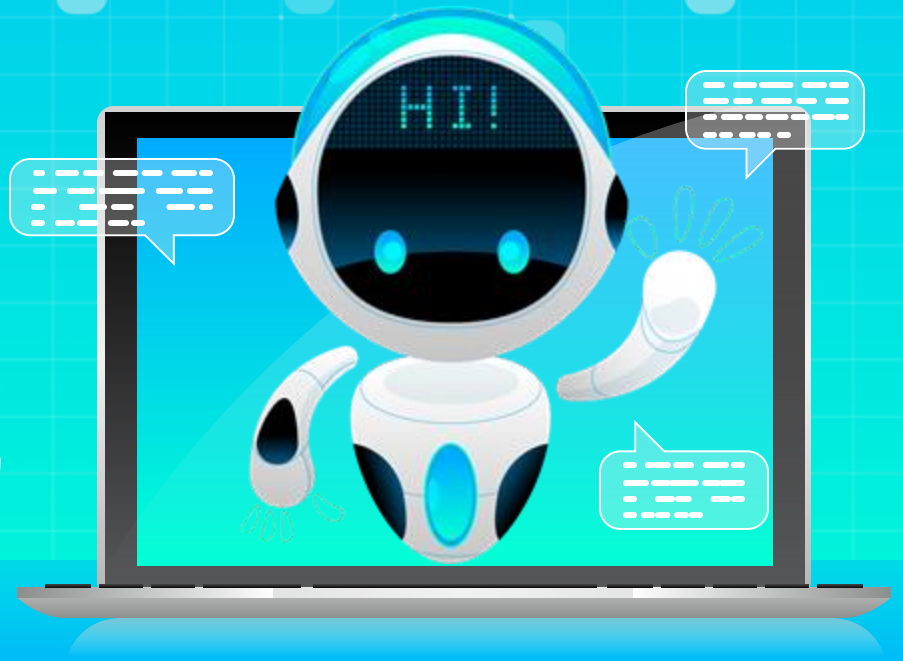
03

AI의 대체

- AI와 대화하며 재미와 오락을 통해 심리적 안정감
- 실시간 응답을 제공하므로 사용자가 즉시 대화를 시작할 수 있음
- 시간과 장소에 구애 받지 않고 대화 가능
- 사람과의 상호작용으로부터 오는 감정 소모 감소

02

서비스 개발 기간



2. 서비스 개발 기간

July 2023

Sun	Mon	Tue	Wed	Thur	Fri	Sat
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29

서비스 기획 및 계획 수립

데이터셋 수집

프론트엔드 개발

데이터 전처리 및 모델 설계

August 2023

Sun	Mon	Tue	Wed	Thur	Fri	Sat
30	31	1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

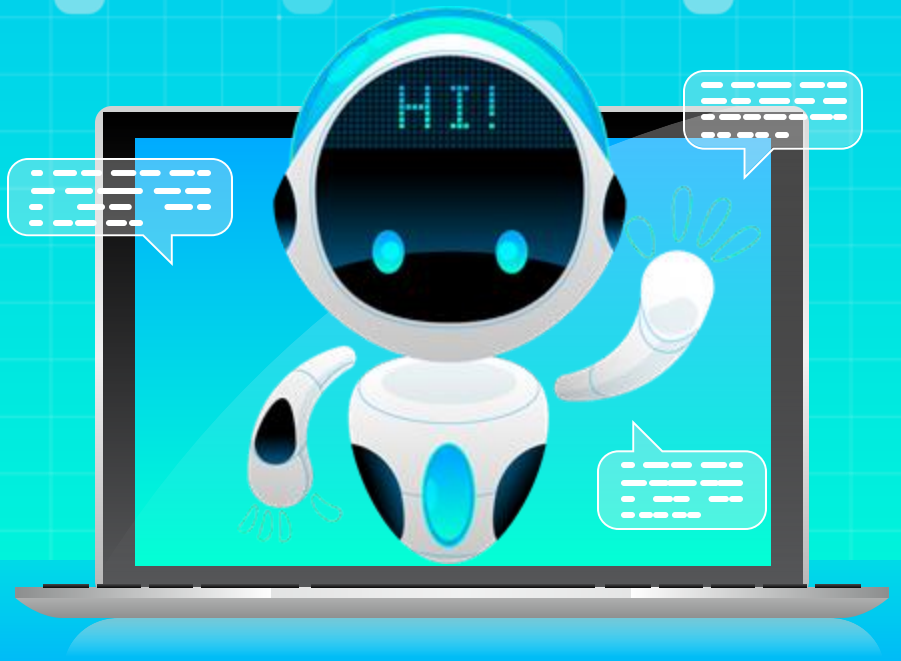
모델 학습 및 파인 튜닝

유지 보수

발표 준비

03

서비스 개발 과정



3.1 데이터셋 수집

모두의 말뭉치 데이터셋

특정 주제 또는 제시 자료로 자유롭게 대화를 나눈 일상 카카오톡 말뭉치 데이터

Data set

T/F 답변 데이터셋

말투의 차별점을 위해 하나의 질문에 대한 T와 F의 답변을 작성하여 약 1800개의 데이터셋 제작

3.2 데이터 전처리 및 학습 데이터 구축

데이터 전처리

전처리

KaKao Dataset을 질문과 대답 형식으로 만들기

- 카카오톡 대화 JSON파일을 파싱하여 **1:1 대화**를 "말하는 사람" : "대화 내용" 형식으로 저장
- 질문과 답변을 "Speaker1"과 "Speaker2"에 저장
- 정규표현식을 통해 특수문자를 제거(@, #, *...)
- 대화의 길이가 2 이상인 데이터만 filter
- 질문에 대한 답변이 나오지 않거나 질문이 연속되는 경우에는 NaN처리
- NaN이 없는 데이터들만 분리하여 csv로 저장

	Speaker_1	Speaker_2
0	1 : 배달 자주 시켜 먹는 편이야?	2 : 아니 주말에는 배달 시켜 먹는 거 같아
1	1 : 주말에만?	NaN
2	1 : 그럼 평일에는 집밥 먹구?	2 : 평일에는 거의 닭가슴살 ㅠㅠ

	Speaker_1	Speaker_2
0	1 : 배달 자주 시켜 먹는 편이야?	2 : 아니 주말에는 배달 시켜 먹는 거 같아
2	1 : 그럼 평일에는 집밥 먹구?	2 : 평일에는 거의 닭가슴살 ㅠㅠ
3	1 : 아 ㅠㅠ 그래도 점심은 밥 먹제?	2 : 어 점심은 일반식 먹지 하하
4	1 : 오늘 점심은 뭐 나와?	2 : 아직 안 정했어 하하
6	1 : 키키 니는 뭐가 먹고 싶어?	2 : 우동 먹고 싶다 하하
...	...	...
499156	1 : 그치 거짓말 조금 보태면 10번 넘게 본 것 같아	2 : 키키 나도 한번 시작하면 끝까지 봐야 해
499157	1 : 그럼 하루 엄청 금방 가잖아 키키 순삭 키키	2 : 맞아 이렇게 끝나기엔 뭔가 아쉬워
499158	1 : 근데 지금 다시 찍으려면 너무 늦었잖아	2 : 그러니깐 ㅠ 이제 해리 자식들로 시작해야지
499159	1 : 근데 해리만큼 잘 할 수 있을까?	2 : 진짜 그 3인방 케미가 최고였는데
499160	1 : 그치그치 완전 최고였는데 ㅜㅠ	2 : 응 다시는 안 나올 조합이야

356030 rows × 2 columns

3.2 데이터 전처리 및 학습 데이터 구축

데이터 전처리

토큰화

- ‘키키’, ‘하하’ 등을 ‘ㅋㅋ’, ‘ㅎㅎ’로 변경
- KoBert Tokenizer을 이용하여 Speaker1과 Speaker2의 문장을 모두 토큰화

	Speaker_1	Speaker_2
0	1 : 배달 자주 시켜 먹는 편이야?	2 : 아니 주말에는 배달 시켜 먹는 거 같아
1	1 : 그럼 평일에는 집밥 먹구?	2 : 평일에는 거의 닭가슴살 ㅠㅠ
2	1 : 아 ㅠㅠ 그래도 점심은 밥 먹제?	2 : 어 점심은 일반식 먹지 하하
3	1 : 오늘 점심은 뭐 나와?	2 : 아직 안 정했어 하하
4	1 : 키키 니는 뭐가 먹고 싶어?	2 : 우동 먹고 싶다 하하

	Speaker_1	Speaker_2
0	배달 자주 시켜 먹는 편이야?	아니 주말에는 배달 시켜 먹는 거 같아
1	그럼 평일에는 집밥 먹구?	평일에는 거의 닭가슴살 ㅠㅠ
2	아 ㅠㅠ 그래도 점심은 밥 먹제?	어 점심은 일반식 먹지 ㅎㅎ
3	오늘 점심은 뭐 나와?	아직 안 정했어 ㅎㅎ
4	ㅋㅋ 니는 뭐가 먹고 싶어?	우동 먹고 싶다 ㅎㅎ

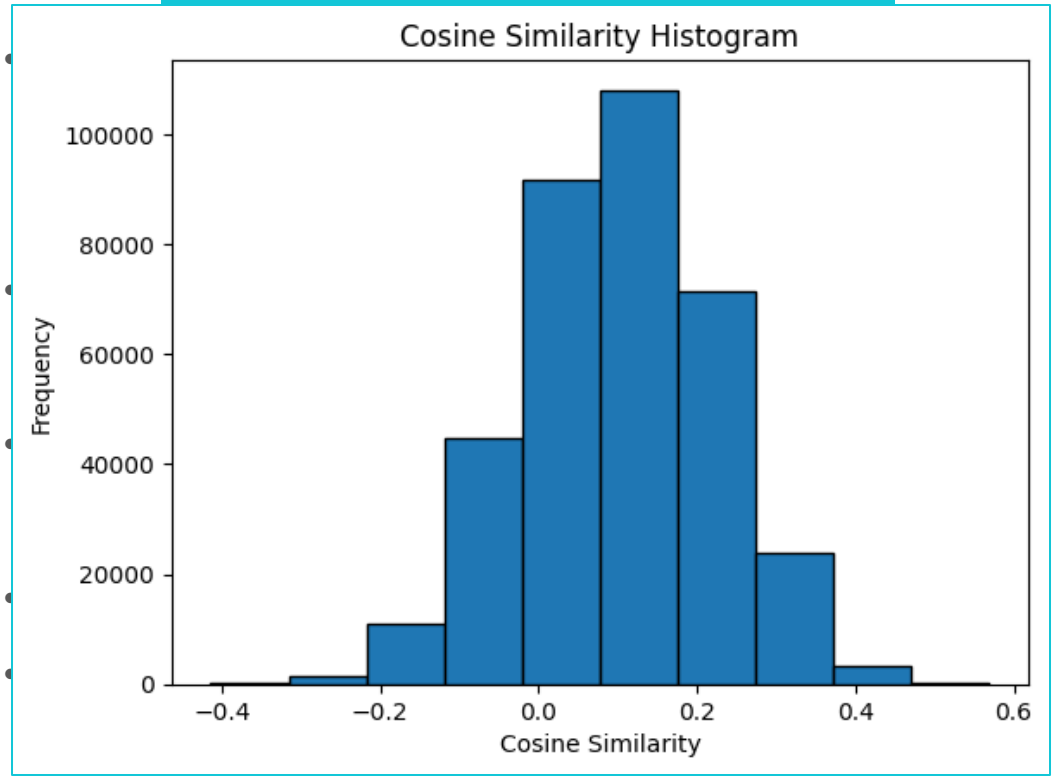
3.2 데이터 전처리 및 학습 데이터 구축

데이터 전처리

임베딩

- Word2Vec 모델을 사용하여 단어 임베딩
- 벡터의 차원은 100차원으로 표현
- 주변 단어를 5개로 설정하여 context를 학습
- 최소 1번 이상 등장하는 단어들을 학습에 사용
- 4개의 쓰레드를 사용하여 학습을 빠르게 진행
- 문장에서 단어들의 벡터를 구해 합산하고 평균을 구함

코사인 유사도 계산



3.3 모델 구축

GPT2

- OpenAI GPT2 논문을 참고
- 자연어 처리 모델로, 입력 받은 텍스트를 기반으로 이어지는 텍스트를 생성하는 모델
- 사전학습과 파인튜닝의 결합으로 만들어진 모델들 중 성능이 가장 우수하지만 여전히 지도학습이 필요
- 충분한 대규모 데이터셋이 없는 경우에는 성능이 떨어질 수 있음
- 과소적합된 모델로 파인튜닝을 통해 원하는 자연어처리 Task에 맞게 사용하는 것 권장
- GPT-2는 영어를 기반으로 한 모델이기 때문에 한국어로 문장을 생성 성능은 상대적으로 정교하지 않음

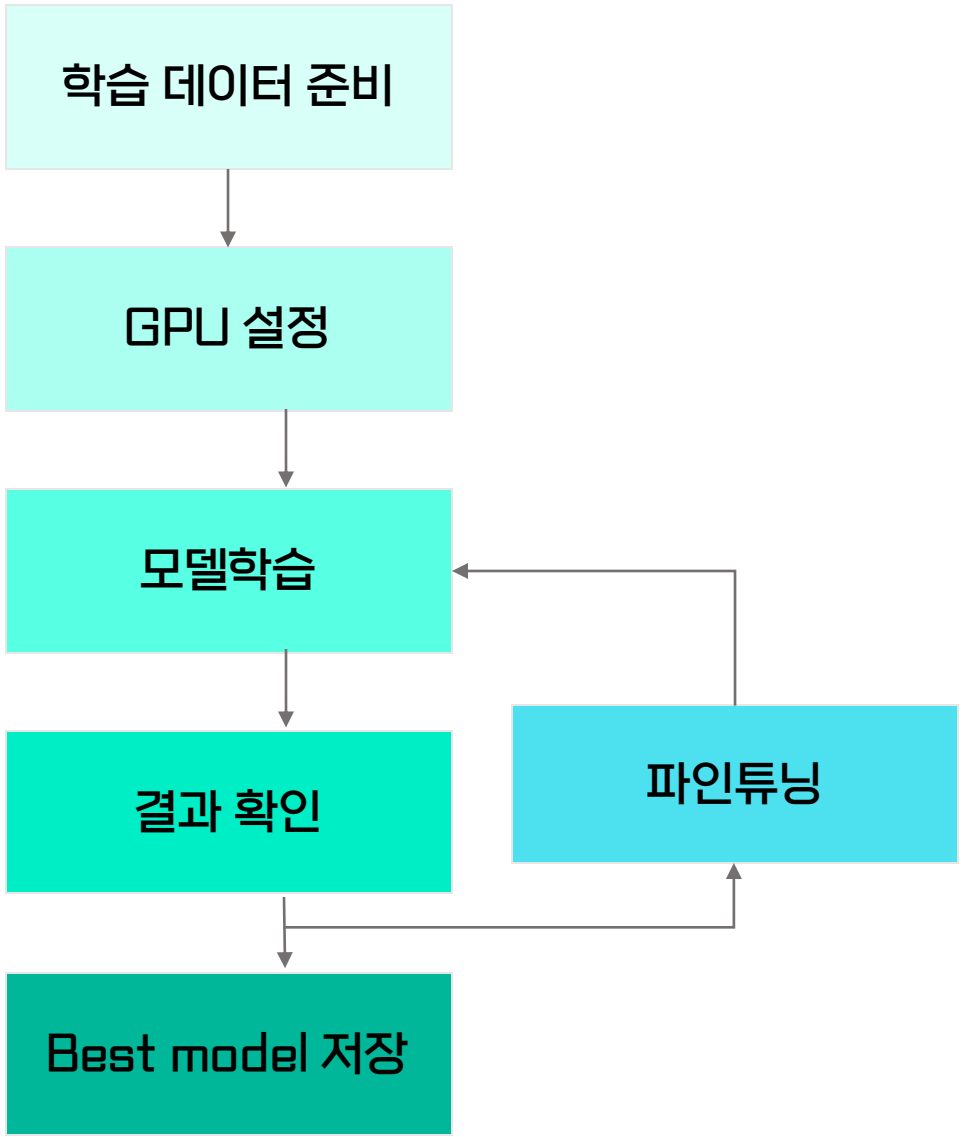
〉 한국어 성능 한계 개선을 위해 개발된 모델이 KoGPT2

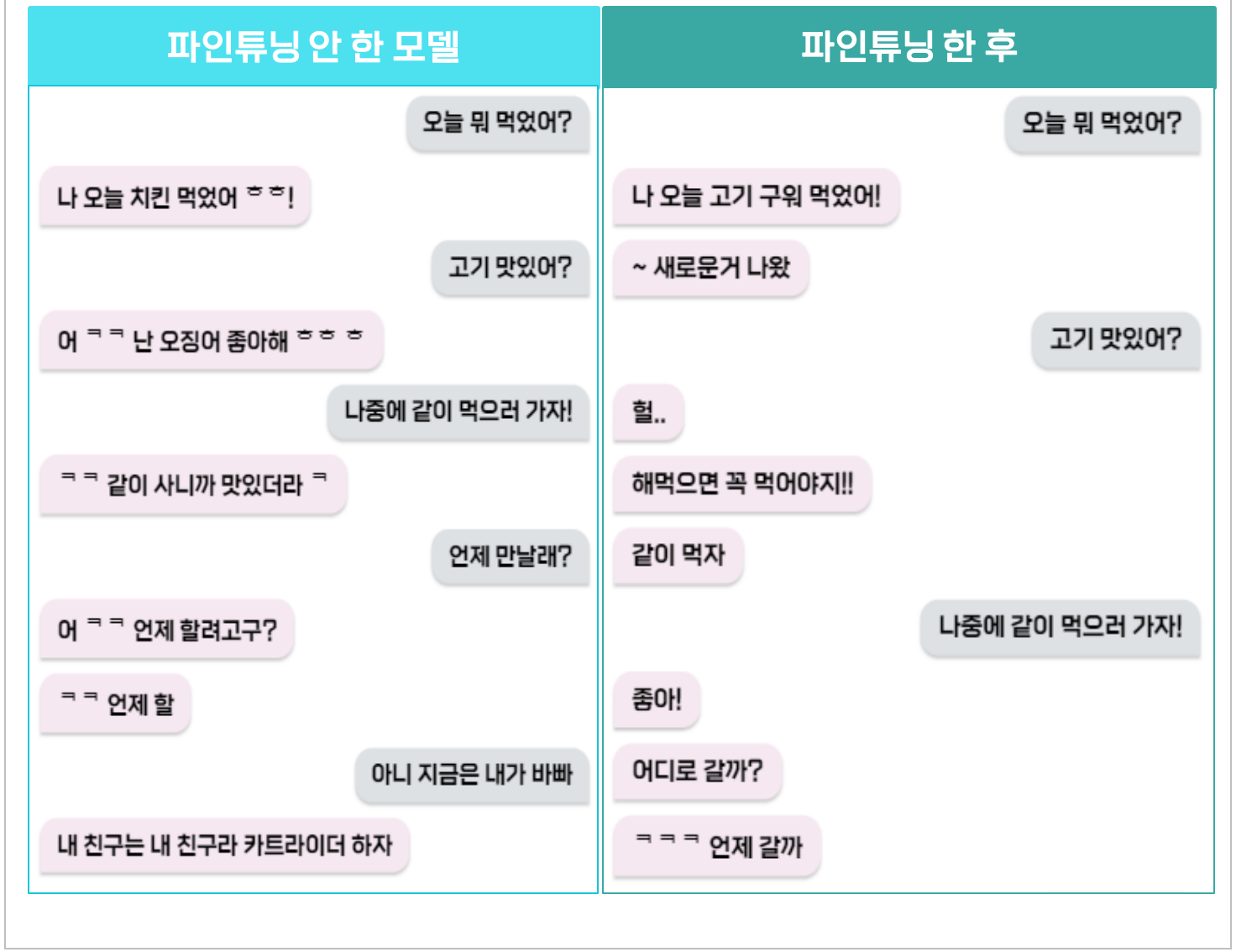
KoGPT 2.0

KoGPT2.0

- KoGPT2.0에서는 문맥을 유지하며 다중 문장 생성에 최적화 되어 의미적으로 연관된 문단을 생성할 수 있음
- 부족한 한국어 성능을 극복하기 위해 40GB 이상의 텍스트로 학습된 모델
- Tokenizer은 Character BPE tokenizer
- 사전 크기는 51,200
- Attention Is All You Need에서 제시한 인코더+디코더 구조에서 인코더 블록을 제거하고 디코더 블록만 사용한 모델

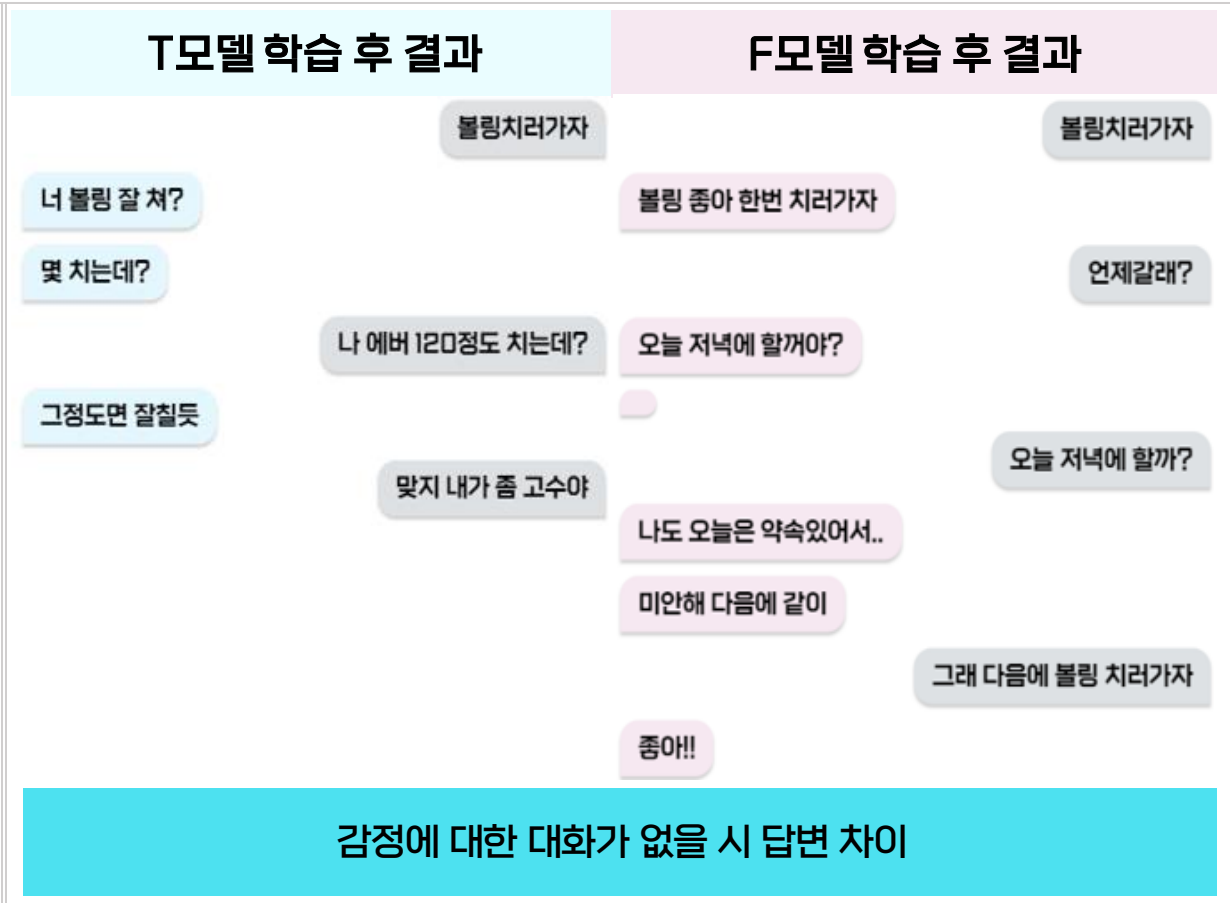
3.4 모델 학습 및 검증


학습 데이터 준비
GPU 설정
모델학습
결과 확인
Best model 저장
파인튜닝


파인튜닝 안 한 모델
오늘 뭐 먹었어?
나 오늘 치킨 먹었어 ㅎㅎ!
고기 맛있어?
어 ㅋㅋ 난 오징어 좋아해 ㅎㅎ ㅎ
나중에 같이 먹으러 가자!
ㅋㅋ 같이 사니까 맛있더라 ㅋ
언제 만날래?
어 ㅋㅋ 언제 할려고구?
ㅋㅋ 언제 할
아니 지금은 내가 바빠
내 친구는 내 친구라 카트라이더 하자
파인튜닝 한 후
오늘 뭐 먹었어?
나 오늘 고기 구워 먹었어!
~ 새로운거 나왔
고기 맛있어?
헐..
해먹으면 꼭 먹어야지!!
같이 먹자
나중에 같이 먹으러 가자!
좋아!
어디로 갈까?
ㅋㅋㅋ 언제 갈까

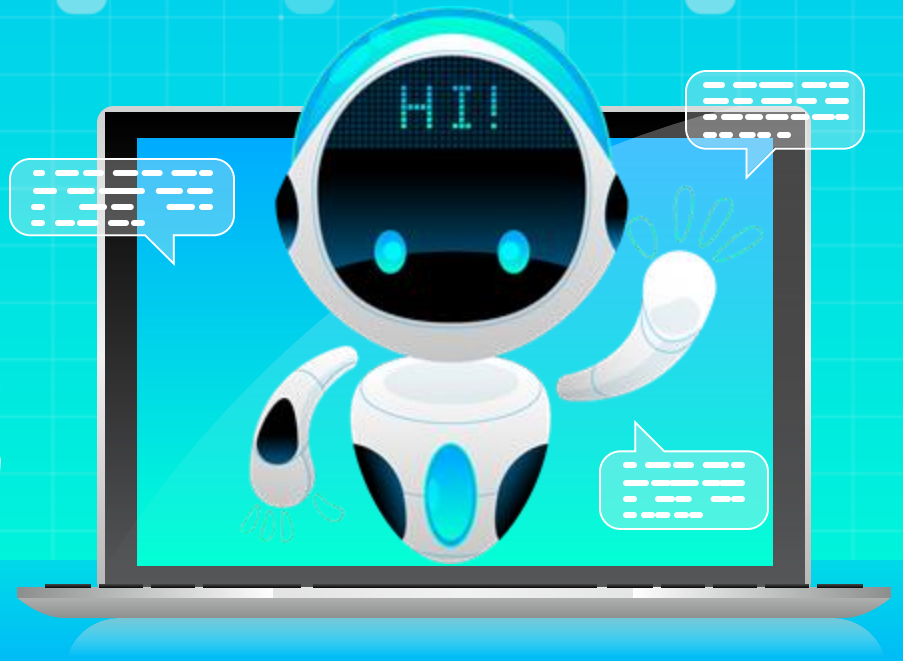
3.4 모델 학습 및 검증

테스트 결과



04

서비스 활용 방안



4. 서비스 활용 방안

MBTI Talk

다기능 AI 메신저

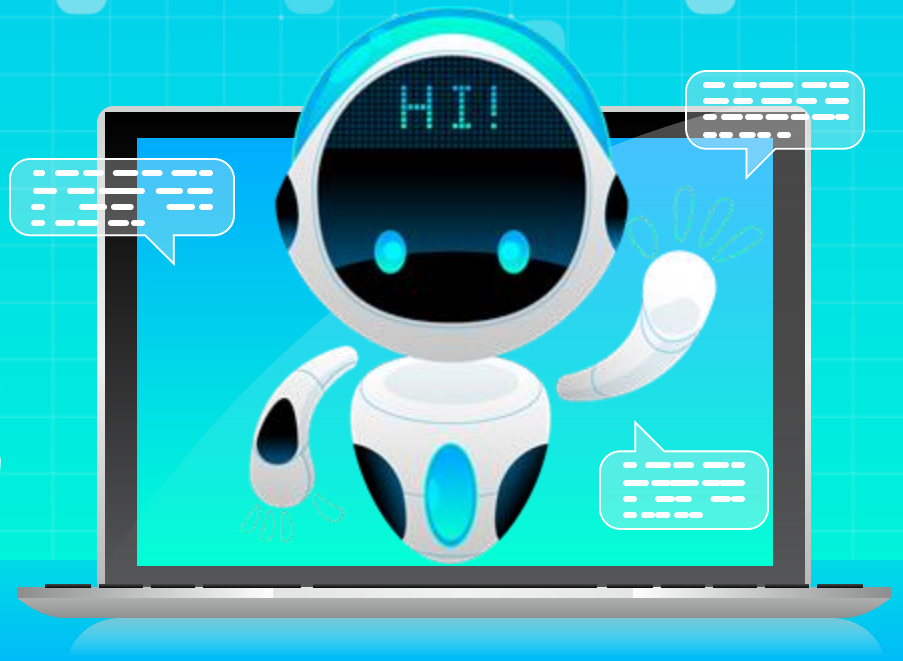
일상 대화 한계를 벗어나 정보를 제공, 제품 추천,
심리 상담 등 추가 학습을 통해
사용자 맞춤 서비스를 제공하는 메신저 APP

카카오톡 채널

일상에서 매일 사용하는 메신저 앱인
'카카오톡'의 채널로 등록하여 범용성을 확대

05

개발 툴



5. 개발 툴

AI

HUGGING FACE

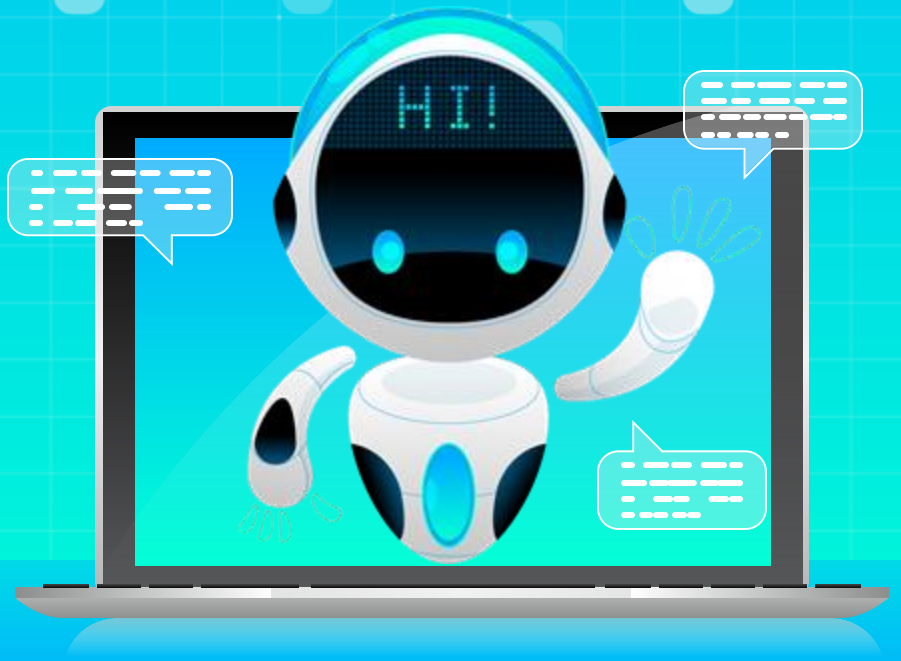
APP

SERVER

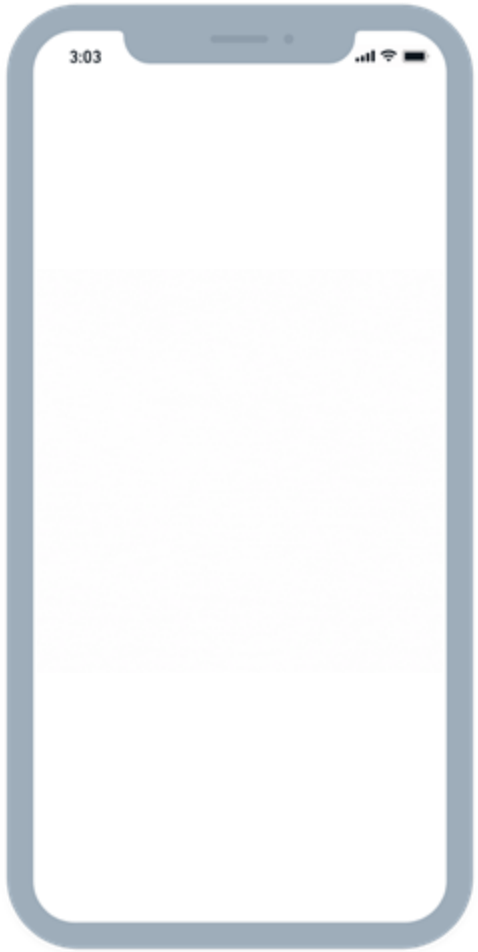
POSTMAN

06

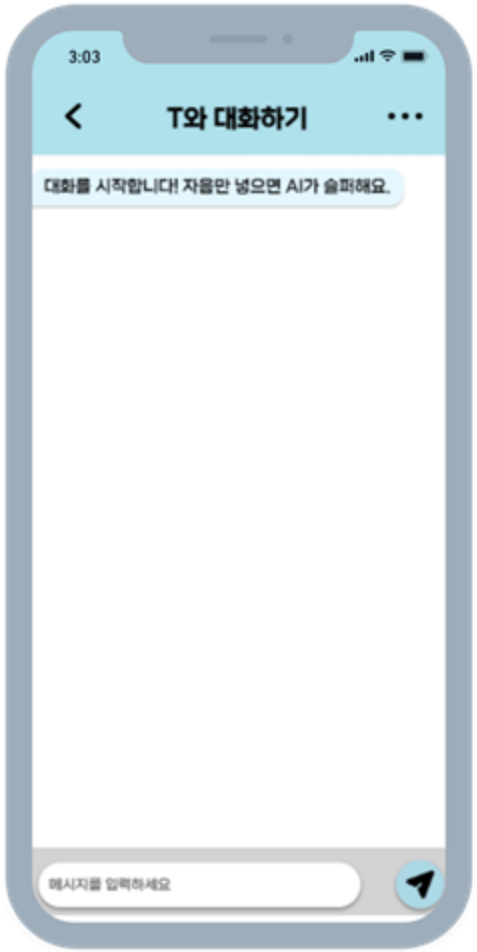
서비스 시연

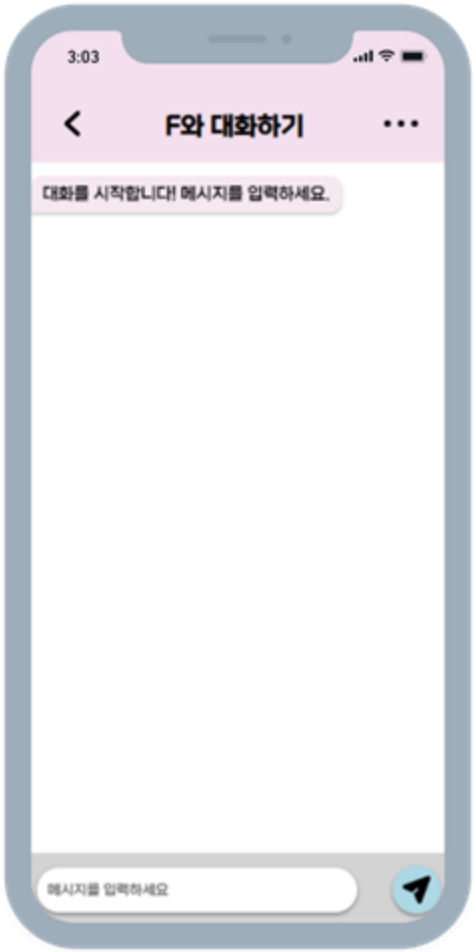


6. 서비스 시연


3:03
Made with Whimsical

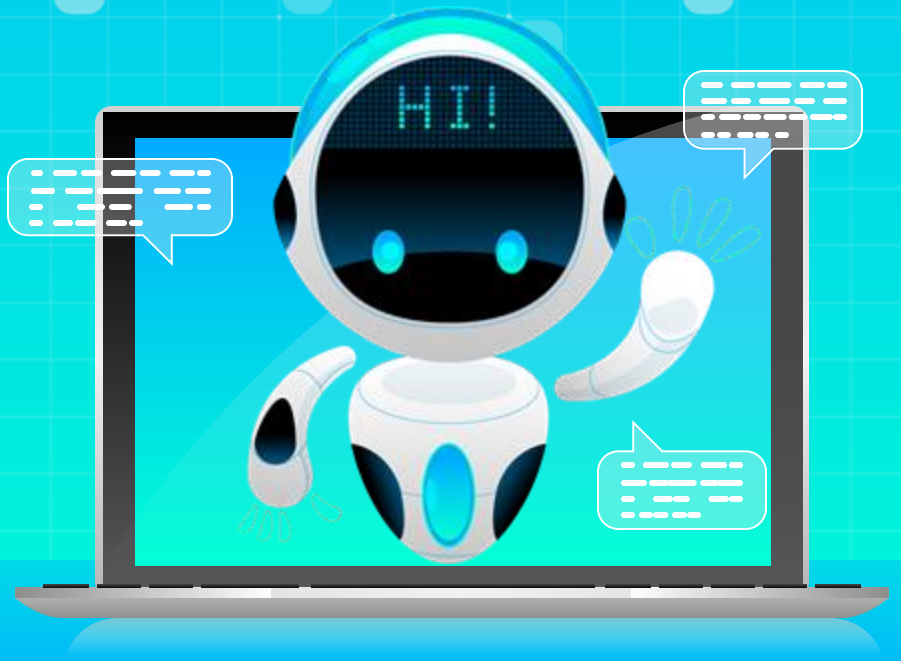

3:03
MBTI Talk
MBTI 대화성향을 기반으로 대화할 수 있어요!
나는 대화할 때
T
Thinking
사고형
F
Feeling
감정형
내 대화성향 간단 테스트하기
Made with Whimsical


3:03
T와 대화하기
대화를 시작합니다! 자음만 넣으면 AI가 술퍼해요.
메시지를 입력하세요
Made with Whimsical


3:03
F와 대화하기
대화를 시작합니다! 메시지를 입력하세요.
메시지를 입력하세요
Made with Whimsical

07

트러블 슈팅



7. 트러블 슈팅

1. MBTI의 T(사고형)과 F(감정형)에 대한 답변이 비슷했다.

MBTI가 T이고 F인 사람들의 답변을 직접 만들어 구별 가능한 답변을 생성

2. 사용자의 질문에 대한 답변이 적절하지 않았다.

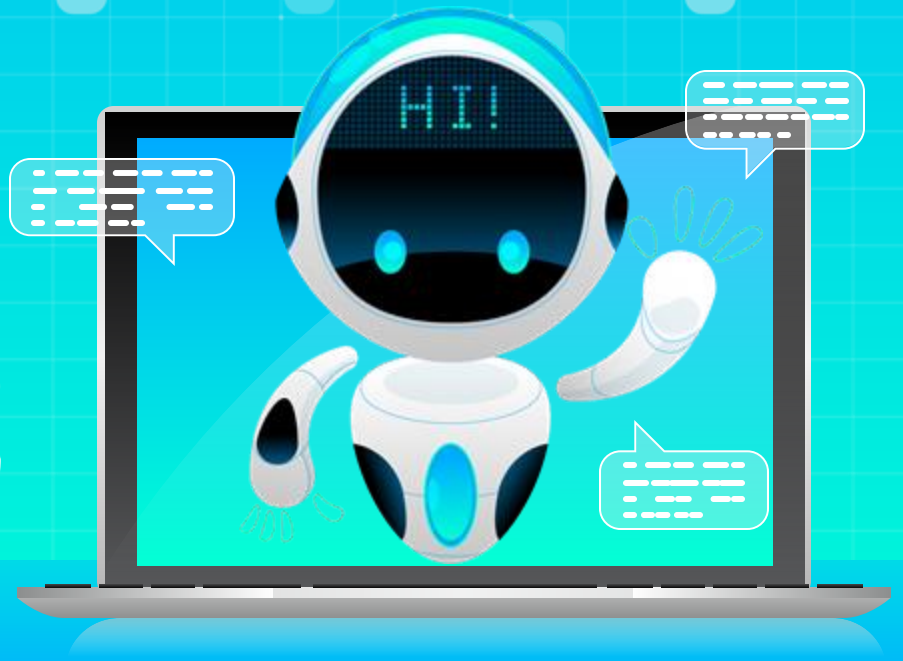
반복적으로 파인 튜닝을 통해 질문에 대한 답변의 정확도를 확보

3. 사용자의 질문에 대해 반복적으로 답변이 출력되었다.

Input값을 토큰화 하여 모델 예측값에서 중복되는 경우 제거하여 출력하도록 함

08

참고문헌



8. 참고 문헌

1. [Open AI] Language Models are Unsupervised Multitask Learners
https://cdn.openai.com/better-languagemodels/language_models_are_unsupervised_multitask_learners.pdf

2. 위키백과 GPT2
https://ko.wikipedia.org/wiki/GPT-2

3. skt-ai KoGPT2
https://github.com/SKT-AI/KoGPT2
https://sktelecom.github.io/project/kogpt2/
https://huggingface.co/skt/kogpt2-base-v2

4. KoGPT2 사용후기
https://medium.com/ai-networkkr/ai-%EB%AA%A8%EB%8D%B8-%ED%83%90%ED%97%98%EA%B8%B0-7-%ED%95%9C%EA%B8%80-%EB%B2%84%EC%A0%84%EC%9D%98-gpt-2-f7317e6499f9

5. 자연어 처리
https://wikidocs.net/book/2155

6. T/F 테스트 모음
https://m.blog.naver.com/lightsmeup/222440353704

THANK YOU