AI 대화형 Chatbot 모델 개발

LLM 모델 사용한 챗봇 서비스

자연어 처리 # AI # LLM

TEAM DBDBDeep.v2
김유진, 이수현, 조서현

고용노동부 직업훈련포털+ HRD-Net 우수훈련기관 서울디지털인재개발원 케이디지털교육센터

Project process

Program 순서도

- 주제 선정
- Dataset Information

주제 선정 & Dataset

Data Preprocessing

- DataFrame 변경
- Dataset 증강

Tokenizer

- skt
- Kykim
- kakao
- Edentns
- LDCC
- RAG

Modeling & Power BI

Loss 값 그래프

Fine Tuning & QLoRA

Chatbot (Streamlit)

- Streamlit
- Flask
- Docker

서비스화

CONTENTS

Tool & Version

python --version → Python 3.12.1

Streamlit --version → Streamilt, version 1.31.0

VSCodeUserSetup-x64-1.85.1.exe

Pycharm-community-2023.3.2.exe

colab → resource 할당 안됨

Kubeflow →

Runpod → 비용이 저렴함, 학습시간이 빠름

1* L40(48GB VRAM 58 GB RAM 16 vCPU) $1.14/hr

Colab Pro = (14800 * 4)
= 58000 (원)

Runpod = 34000 (원)

MSOffice = 89000 (원) / 1 year

Hugging Face 주소

https://huggingface.co/

GitHub 주소

https://github.com/suhyun0115/LLM

주제선정 및 Dataset

주제선정

[배경]

- 도배 하자와 관련된 다양한 질문과 상황을 제공하고,
  이에 대한 정확하고 신속한 응답을 제공하는 AI 모델을 개발하는 것을 목표

- 이는 실제 현장에서 발생할 수 있는 복잡한 상황에 대응하고,
  고객의 문의에 신속하고 정확하게 답변할 수 있는 시스템을 구축하는 데
  중요한 역할을 할 것

[Data] https://dacon.io/competitions/official/236216/data

Dataset

[Dataset Info]

- train.csv [파일]
  - id : 질문 – 답변 (QA) 샘플 고유 번호
  - 질문_1, 질문_2 : 샘플 별 동일한 내용으로 구성된 질문 2개
  - category : 질문 – 답변 (QA) 샘플의 도메인 세부 분야
  - 답변_1, 답변_2, 답변_3, 답변_4, 답변_5 :
    샘플 별 질문에 대한 동일한 답변 Reference 5개

- test.csv [파일]
  - id : 평가 질문 샘플 고유 번호
  - 질문 : 평가 샘플의 질의 내용

- sample_submission.csv [파일] – 제출양식
  - id : 평가 질문 샘플 고유 번호
  - vec_0, vec_1, ..., vec_511 :
    생성된 답변을 512 차원의 Embedding Vector로 표현된 결과

Data Preprocessing

DataFrame 변경

- train.csv 파일 dataframe 변경하기

(질문_1, 질문_2)

(답변_1, 답변 _2, 답변_3, 답변_4, 답변_5)

	id	질문_1	질문_2	category	답변_1	답변_2	답변_3	답변_4	답변_5
0	TRAIN_000	면진장치가 뭐야?	면진장치에 사용되는 주요 기술은 무엇인가요?	건축구조	면진장치란 지반에서 오는 진동 에너지를 흡수하여 건물에 주는 진동을 풀여주는 진동 ...	면진장치란 건물의 지반에서 발생하는 진동 에너지를 흡수하여 건물을 보호하고, 진동을...	면진장치란 지반으로부터 발생하는 진동 에너지를 흡수하여 건물에 전달되는 진동을 풀여...	면진장치는 건물의 지반으로부터 오는 진동 에너지를 흡수하여 건물에 전달되는 진동을 ...	면진장치는 건물에 오는 지반 진동의 영향을 최대한으로 흡수하여 건물에 전달되는 진동...
1	TRAIN_001	내진설계의 종류좀 알려줘	내진설계에는 어떤 종류가 있는지 자세히 알려주실 수 있나요?	건축구조	내진 설계의 종류로 내진구조, 제진구조, 면진구조가 있습니다.	내진설계에는 내진구조, 제진구조, 면진구조가 있습니다. 내진구조는 건물 구조물이지...	내진설계에는 주로 내진구조, 제진구조, 면진구조의 세 가지 종류가 있습니다. 이들은...	내진설계에는 주로 내진구조, 제진구조, 면진구조가 사용됩니다. 내진구조는 건물 구조...	내진 설계에는 다양한 종류가 있지만, 대표적으로 내진구조, 제진구조, 면진구조가 있...
2	TRAIN_002	철골구조의 장점이 뭐야?	철골구조의 장점을 알려줘?	건축구조	철골구조는 건물의 외벽에는 그다지 하중이 걸리지 않기 때문에 고층 건물의 건축이 가...	철골구조의 장점은 건물의 외벽에는 그다지 하중이 걸리지 않기 때문에 고층 건물의 건...	철골구조의 장점은 건물의 외벽에 하중이 적게 걸리기 때문에 고층 건물의 건축이 용이...	철골구조의 장점은 건물의 외벽이 하중이 걸리지 않아 공간 활용이 용이하고, 고층 건...	철골구조의 장점은 건물의 외벽에 하중이 크게 걸리지 않아 고층 건물을 건축할 수 있...
3	TRAIN_003	철골철근 콘크리트 구조가 뭐야?	철골철근 콘크리트 구조의 장점과 단점에는 무엇이 있을까요?	건축구조	철근철골콘크리트는 철골과 철근, 그리고 콘크리트를 함께 사용하는 건축 구조입니다. ...	철골철근콘크리트 구조는 건축물을 지탱하는 주요 구조물인 철골과 철근, 그리고 콘크리...	철골철근 콘크리트 구조는 건축물을 지탱하기 위한 구조물에서 일반적으로 사용되는 방식...	철골철근콘크리트 구조는 철골과 철근, 그리고 콘크리트를 함께 사용하여 만들어지는 건...	철골철근 콘크리트 구조는 강철 골조와 강철 철근, 그리고 콘크리트를 함께 사용하여 ...
4	TRAIN_004	철골구조는 어떤 방식이 있어?	철골구조의 다양한 방식이 무엇인가요?	건축구조	철골구조는 일반철골구조와 경량철골구조가 있습니다.	철골구조는 일반철골구조와 경량철골구조가 있습니다. 일반철골구조는 주로 대형 건물이나...	철골구조는 주로 일반철골구조와 경량철골구조로 나뉘어집니다. 이들은 건축 시스템에 따...	철골구조는 주로 일반철골구조와 경량철골구조로 구분됩니다. 이외에도 최근에는 고층 건...	철골구조는 일반철골구조와 경량철골구조 두 가지 방식이 주로 사용됩니다. 일반철골구조...

각 <usr>사용자 입력과 <sys> 챗봇 답변
→ 한 쌍으로 매칭되도록 DataFrame 변경

[변경]
TRAIN_000 질문_1 – 답변_1
TRAIN_000 질문_1 – 답변_2
TRAIN_000 질문_1 – 답변_3
TRAIN_000 질문_1 – 답변_4
TRAIN_000 질문_1 – 답변_5
TRAIN_000 질문_2 – 답변_1
TRAIN_000 질문_2 – 답변_2
TRAIN_000 질문_2 – 답변_3
TRAIN_000 질문_2 – 답변_4
TRAIN_000 질문_2 – 답변_5
…

Data Preprocessing

DataSet 증강

새로운 카테고리 생성 후 질문과 답변 5개 생성 → concat( ) 후 csv file 저장

NO	카테고리	질문1	질문2	답변1
1	가구수정	벽지를 붙일 때 정확한 위치는 어떻게 찾지?	벽지를 붙일 때 정확한 위치를 어떻게 찾나요?	벽지를 정확한 위치에 붙이기 위해선 먼저 벽의 수평과 수직을 확인하고, 시작점을 정한 후에 롤러나 정밀도구를 사용하여 벽지의 위치를 조정하십시오.
2	가구수정	벽지 위치를 잘못 붙였을 때 어떻게 고치지?	벽지 위치를 잘못 붙였을 때 어떻게 수정하나요?	벽지를 잘못 붙인 경우, 가능한 빨리 벽지를 조심스럽게 떼어내어 재조정하시는 것이 좋습니다. 이미 완전히 마른 경우, 전문가의 도움을 받는 것이 바람직합니다.
3	가구수정	벽지 붙이기 전에 뭐 준비해야 돼?	벽지 붙이기 전에 무엇을 준비해야 하나요?	벽지를 붙이기 전에는 벽의 상태를 점검하고, 필요한 경우 벽을 매끄럽게 하고, 청소하여 먼지와 기름기를 제거해야 합니다.
4	가구수정	벽지 패턴이 제대로 안 맞을 때는 어떡해야 해?	벽지 패턴이 정확하게 맞지 않을 때는 어떻게 해야 하나요?	벽지 패턴이 맞지 않는 경우, 패턴이 시작되는 지점을 정확하게 측정하고, 패턴을 일치시키기 위해 추가적인 조정이 필요할 수 있습니다.
5	가구수정	벽지 붙일 때 가장 중요한 건 뭐라고 생각해?	벽지를 붙일 때 가장 중요한 점은 무엇인가요?	벽지를 붙일 때는 정확한 위치 선정, 패턴의 일치, 공기방울 없이 매끄럽게 붙이는 것이 가장 중요합니다.

NO	카테고리	질문1	질문2	답변1	답변2	답변3	답변4	답변5
1	가구수정	벽지를 붙일 때 정확한 위치는 어떻게 찾지?	벽지를 붙일 때 정확한 위치를 어떻게 찾나요?	벽지를 정확한 위치에 붙이기 위해선 먼저 벽의 수평과 수직을 확인하고, 시작점을 ...	벽지를 붙일 때는 벽지의 패턴이나 디자인이 맞도록 조심스럽게 배치하는 것이 중요합니...	벽지를 부착할 때는 시작점을 정한 후에 벽의 수평과 수직을 확인하는 것이 중요합니다...	벽지를 부착할 때는 먼저 벽의 수평과 수직을 정확히 확인한 후에 시작점을 결정해야 ...	벽지를 붙이기 전에는 레이저 수평계나 수평 앱을 사용하여 벽의 수평과 수직을 정밀하...
2	가구수정	벽지 위치를 잘못 붙였을 때 어떻게 고치지?	벽지 위치를 잘못 붙였을 때 어떻게 수정하나요?	벽지를 잘못 붙인 경우, 가능한 빨리 벽지를 조심스럽게 떼어내어 재조정하시는 것이...	벽지를 잘못 붙인 경우에는 벽지를 조심스럽게 떼어내어 재조정하는 것뿐만 아니라, 벽...	벽지를 재조정할 때는 벽지를 완전히 제거하고, 벽면을 청소하여 먼지와 기름기를 제거...	벽지를 재조정할 때에는 먼저 부착된 벽지를 조심스럽게 제거하는 것이 중요합니다. 벽...	벽지를 재조정하는 과정에서는 벽면의 표면을 깔끔하게 정리하고, 필요한 경우 보정 작...
3	가구수정	벽지 붙이기 전에 뭐 준비해야 돼?	벽지 붙이기 전에 무엇을 준비해야 하나요?	벽지를 붙이기 전에는 벽의 상태를 점검하고, 필요한 경우 벽을 매끄럽게 하고, 청...	벽지를 부착하기 전에는 벽의 표면을 철저히 점검하여 부식, 균열, 또는 기타 결함...	벽지를 부착하기 전에는 벽의 표면을 철저히 검토하여 과거의 부착물이나 다른 불순물이...	벽지를 부착하기 전에는 벽의 표면을 부드럽게 갈거나 수평을 맞추어야 합니다. 벽면의...	벽지를 부착하기 전에는 벽의 표면을 깨끗하게 닦아야 합니다. 벽면에 먼지나 기름기가...
4	가구수정	벽지 패턴이 제대로 안 맞을 때는 어떡해야 해?	벽지 패턴이 정확하게 맞지 않을 때는 어떻게 해야 하나요?	벽지 패턴이 맞지 않는 경우, 패턴이 시작되는 지점을 정확하게 측정하고, 패턴을 ...	벽지 패턴이 맞지 않을 때는 패턴의 일치 여부를 확인하고 추가적인 조정이 필요한 부...	벽지 패턴이 일치하지 않는 경우, 패턴을 조정하는 데에는 주의가 필요합니다. 먼저 ...	벽지 패턴이 맞지 않는 경우, 정확한 패턴 일치를 위해 조정이 필요합니다. 패턴이 ...	벽지 패턴이 맞지 않을 때에는 조정 작업이 필요합니다. 패턴이 시작되는 위치를 정확...
5	가구수정	벽지 붙일 때 가장 중요한 건 뭐라고 생각해?	벽지를 붙일 때 가장 중요한 점은 무엇인가요?	벽지를 붙일 때는 정확한 위치 선정, 패턴의 일치, 공기방울 없이 매끄럽게 붙이는...	벽지를 부착할 때는 패턴의 일치뿐만 아니라, 공기방울이 없이 매끄럽게 붙이는 것도 ...	벽지를 부착할 때는 패턴의 일치와 더불어 공기방울이 생기지 않도록 주의해야 합니다....	벽지를 부착할 때는 정확한 위치를 선정하여 패턴을 일치시키는 것이 중요합니다. 또한...	벽지를 부착할 때는 정확한 위치를 선택하고, 패턴을 일치시키는 것뿐만 아니라 공기방...

```
df = pd.concat([new_df, new_df_df], ignore_index=True)
df.to_csv('/content/drive/MyDrive/DBDBDeep/last_df.csv', index=False)
df.tail()
```

	Question	Answer
7400	벽지면과 벽지면 사이가 벌어졌을 때는 어떤 전문가의 도움이 필요할까?	벽지면과 벽지면 사이가 벌어진 경우, 건축 전문가나 구조 공학자의 도움이 필요할 수...
7401	벽지면과 벽지면 사이가 벌어졌을 때는 어떤 전문가의 도움이 필요할까?	벽지면과 벽지면 사이가 벌어졌을 때는 먼저 집의 구조에 대한 이해가 필요합니다. 구...
7402	벽지면과 벽지면 사이가 벌어졌을 때는 어떤 전문가의 도움이 필요할까?	벽지면과 벽지면 사이가 벌어졌는데 벽체의 수축이나 건축 재료의 변형으로 인한 것으로...
7403	벽지면과 벽지면 사이가 벌어졌을 때는 어떤 전문가의 도움이 필요할까?	벽지면과 벽지면 사이가 벌어졌을 때, 주변 환경에 따라 습기, 온도, 땅의 침하 등...
7404	벽지면과 벽지면 사이가 벌어졌을 때는 어떤 전문가의 도움이 필요할까?	벽지면과 벽지면 사이가 벌어졌을 때, 건축 디자인의 결함 또는 잘못된 시공이 원인일...
7405	벽지면과 벽지면 사이가 벌어졌을 때는 어떤 전문가의 도움이 필요할까요?	벽지면과 벽지면 사이가 벌어진 경우, 건축 전문가나 구조 공학자의 도움이 필요할 수...

•••

6440 → 7430 (dataset 990개 증가)
총 dataset 개수 7430개

"Id" 와 "category" 삭제

토크나이저 사용으로 stop word(불용어) 사용 안함

- 특수문자 또는 공백 등 기준으로 문장 나눔
- 토크나이저는 단어들을 분리된 토큰으로 취급하지 않음
- 스탑워즈를 명시적으로 제거 필요 없음

Model 설명

LLM 모델 성능평가는 총 5가지
추론능력, 상식능력, 언어 이해력, 환각방지능력, 한국어 일반상식능력에 대해서 평가

평가지표	주요내용
추론능력 (ARC)	- (ARC, A12 Reasoning Challenge) AI가 질문에 대한 답변이 얼마나 적합한지를 측정 ※ 초등학교 수준의 과학 질문지로만 구성
상식능력 (HellaSwag)	- (HellaSwag) AI가 짧은 글 및 지시사항에 알맞은 문장을 생성하는지 여부 측정 ※ 인간에게는 사소한 질문이지만, AI에게는 답변하기 어려운 질문지로 구성
언어 이해력 (MMLU)	- (MMLU, Massive Multitask Language Understanding) 방대한 분야의 질문에 대한 답변이 얼마나 정확한지를 측정 ※ 57개 다양한 분야(초등 수학, 역사, 컴퓨터 과학, 법학 등)에 대한 질문지로 구성
환각방지능력 (TruthfulQA)	- (TruthfulQA) AI가 생성한 답변이 얼마나 진실한지 측정 ※ 인간이 잘못 인지 or 거짓으로 대답할 수 있는 질문지로 구성
한국어 상식생성능력	- (Korean-CommonGEN-V2) AI가 주어진 조건의 질문에 대한 답변이 한국어 사용자라면 보유하고 있을 일반 상식에 부합하는지 여부 측정 ※ 역사 왜곡, 환각오류, 형태소 부착 오류, 불규칙 활용 오류, 혐오 표현 등에 대한 광범위한 유형을 포함한 설문지로 구성

[AIHub] https://www.aihub.or.kr/leaderboard/view.do?currMenu=500&topMenu=102

Modeling

Model 설명

특징	GPT	BERT	LLAMA
목표	다음 단어 예측 및 생성	양방향 문맥 이해 및 단어 표현 학습	다음 문장 예측을 중심으로 문장 생성
학습 방식	단방향 Transformer를 사용하여 미세 조정(fine-tuning)	양방향 Transformer를 사용하여 사전 학습(pre-training)	양방향 Transformer를 사용하여 사전 학습(pre-training)
활용	- 문장 생성 - 기계 번역 - 요약 - 질의 응답 등	- 문장 분류 - 질의 응답 - 단어 임베딩 등	- 문장 생성 - 기계 번역 - 요약 - 질의 응답 등
주요 특징	다음 단어 예측	문장 내 단어 의미 이해 문맥 파악	다음 문장을 예측하는 방식으로 문장 생성

- **GPT1: 언어 이해를 개선하기 위한 생성적 사전 학습 모델**
- **GPT2: 언어 모델은 비지도 학습 다중 작업 학습자들**
- **GPT3: 언어 모델은 소량의 학습 데이터로도 학습하는 모델**

- **GPT는 생산적 언어 모델로 다음 언어를 예측하는 것에 중점으로 함**
- **BERT는 양방향으로 문맥을 이해하여 문장 내 단어의 의미를 파악함**
- **LLAMA는 지식 그래프를 활용하여 다음 문장을 예측하는 생성적 언어 모델**

Model Loss 값 측정 기준

Loss 값의 측정 기준은 모델이 예측한 출력과 실제 레이블 간의 차이를 나타내는 손실

Training Dataset

주어진 데이터셋을 학습하는 과정에서 각 배치를 모델에 입력

손실 계산

해당 배치에 대한 출력과 레이블을 비교하여 손실 계산

모델이 예측한 출력과 실제 레이블 간의 차이를 나타냄
차이가 적을수록 손실이 작음

loss.mean( )

코드에서 loss.mean( )을 통해 배치 손실 평균 계산

각 Epoch의 전체 손실 계산함

Loss 출력

최종적으로 Epoch 당 평균 손실이 출력함

Early stopping

학습 중 최적의 모델을 저장하고 Early stopping 조건을 사용하여 학습을 조기 종료

Modeling (결론)

Model 스펙

	SKT	*kykim*	*Kakao*	*Edentns*	*LDCC*	*RAG*
사용 스펙	Colab T4 GPU	Colab T4 GPU	Colab A100	KubeFlow Runpod	KubeFlow Runpod	KubeFlow
Tuning	Fine Tuning	Fine Tuning	Fine Tuning	QLoRA	QLoRA	-
학습 용량	RAM : 0.9 GB 그래픽 : 15 MB	RAM : 1 GB 그래픽 : 30 MB	RAM : 25 GB 그래픽 : 16 GB	RAM : 23 GB 그래픽 : 8 GB	RAM : 23 GB 그래픽 : 8 GB	RAM : 16 GB 그래픽 : 5 GB
학습 시간 **(1 Epoch 당)**	00 : 01 : 10~11	00 : 08 : 11~12	00 : 27 : 43~51	06 : 04 : 40 ~ 06 : 05 : 44	01 : 26 : 15 ~ 01 : 26 : 27	-
총 학습 시간	00 : 40 : 15	01 : 06 : 01	04 : 37 : 11	12 : 10 : 24	11 : 30 : 39	-

Fine Tuning : 모든 하이퍼파라미터를 조정
LoRA : 사전 훈련된 모델에 대해 언어 관련 작업 미세 조정을 수행
QLoRA : LoRA + 양자화 -> 성능 하락 없이 메모리를 절약

Modeling

01 skt/kogpt2-base-v2

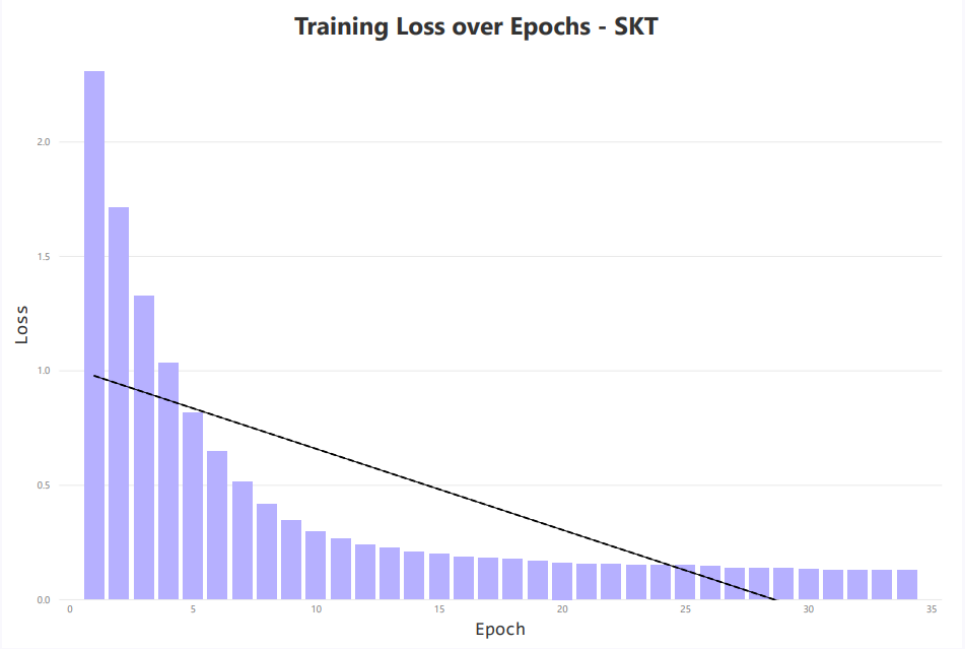
- 한국어 자연어 처리를 위해 개발된 GPT-2 모델
- SK Telecom & KAIST 공동 개발
  → KoGPT2모델 개선 버전
- 한국어 표현을 더 자연스럽게 만들기 위해 사용
- 문맥을 유지하면서 더 많은 문장을 생성할 수 있게 됨
  → 학습 Dataset 및 모델 크기 UP↑
- 경제적 활용 가능한 크기 유지하면서 성능 향상↑

```
return_answer_by_chatbot('도배할때 주의사항이 뭐야')
```

```
'도배 할 때는 벽의 표면을 철저히 점검하여 부식, 균열, 또는 기타 결함이 있는지 확인해야 합니다. 필요한 경우 해당 결함을 보정하고, 결함을 정확히 파악하여 적절한 조치를 취해야 합니다.'
```

```
return_answer_by_chatbot('도배비용 알려줘')
```

```
'일반적으로 도배 견적은 (도배평수/5*도배지가격)+인건비+부자재로 산출합니다.'
```



02 kykim/gpt3-kor-small_based_on_gpt2

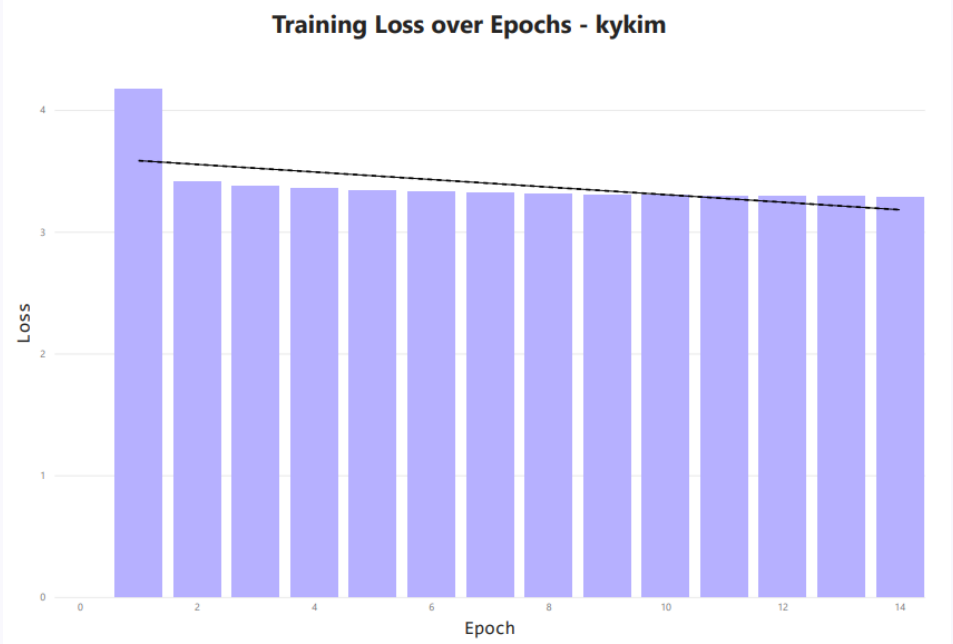
```
1 return_answer_by_chatbot('도배지에 얼룩이 생기는 다양한 원인들에 대해서 자세히 알려주세요.')
```

도배지의 녹은 흔적이 벽지에 남아 있을 경우, 이는 접착제나 페인트를 사용하여 복원이 가능합니다. 그러나 이러한 방법으로는 한계가 있습니다.

```
1 return_answer_by_chatbot('준불연재료는 무엇인가요? 그리고 유성페인트를 사용하는 것에 대한 부작용이 있을까요?')
```

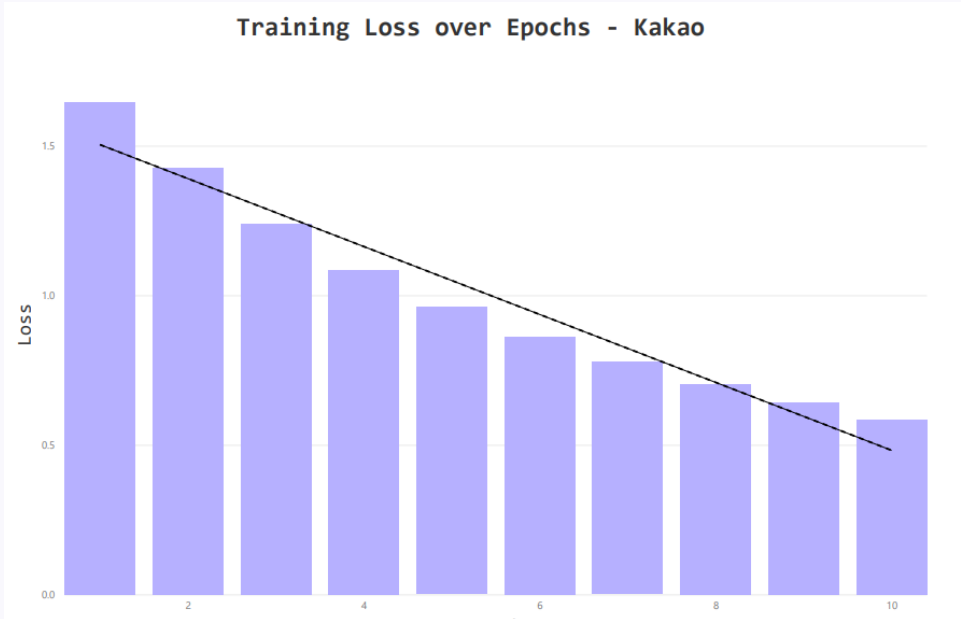
도배에 사용된 풀의 접착력이 약해 벽지가 쉽게 떨어질 수 있습니다. 또한, 이러한 현상은 습기와 온도 변화에 민감하게 반응하여 발생할 수도 있으므로 주의가 필요합니다.

- Bert base model for Korea 한국어에 대한 BERT의 기본 모델
- 한국어 자연어 처리를 위해 만들어졌으며, BERT 아키텍처를 기반으로 함
- 70GB 크기의 한국어 텍스트 데이터셋을 사용하여 학습됨
- 텍스트를 위해 42000개의 소문자로 변환된 하위 단어 사용



03 kakaobrain/kogpt

- GPT-3 모델의 한국어 특화 버전인 'KoGPT'를 오픈소스 공개
- 한국어를 사전적, 문맥적으로 이해하여 다양한 언어 과제를 수행함
  → 감성 분석부터 글쓰기까지 다양한 언어 작업을 자동화 할 수 있음
- 동일한 문장들을 GPT-3와 GPT-2에 넣었을 때, GPT-3가 전반적으로 더 그럴듯한 문장을 생성함




```
1 user_text='도배하는 법 알려줘'
```

```
1 return_answer_by_chatbot(user_text)
```

```
'? 나도 잘 몰라. '
```

```
1 user_text='벽에 곰팡이가 폈어'
```

```
1 return_answer_by_chatbot(user_text)
```

```
" "어? 어디야?" "여기야, 내가 사는 집이야. 그런데, 이 사람들, 누구지? 이 집에 사는 사람인가?" "어, 그래. 그런데, 누구지?" "누구냐고 묻지마. 그냥 아는 사람이야. 내가 좋아하는 사람이고.
룩으로 일관했다. 나는 그녀가 지금 나에 대해 뭔가 숨기고 있다는 생각이 들었다. "이 사람, 어디서 만났죠?" 내가 물었다. "어, 내가 전에 살던 아파트 근처였어."'
```

04 Edentns/DataVortexS-10.7B-dpo-v1.11

- AI Hub 한국어 LLM 리더보드에서 1위로 롯데정보통신 LDCC/LDCC-SOLAR-10.7B 기반
- 짧은 글 및 지시사항 생성에 뛰어남
- 적용한 매개변수는 10.9억개
- 상업용 목적으로 사용 불가

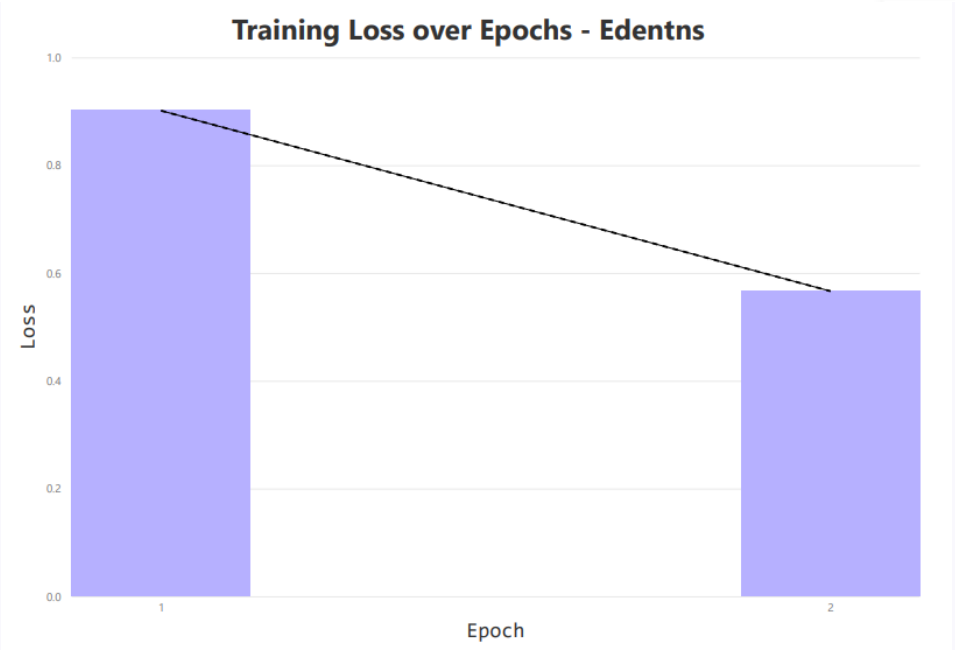
```
user_text ='도배하는 법 알려줘'
```

```
return_answer_by_chatbot(user_text)
```

```
'도배 방법 은 1. 기존 벽지를 제거하고 벽면을 평탄하게 만든 후, 2. 벽면에 초배지를 부착하고 마르 게 한 후 3. 실배지를 벽면에 부착합니다 .'
```

```
return_answer_by_chatbot('좀 더 자세히 알려줘')
```

```
'벽지를 붙 일 때는 벽 의 모서리와 수직을 먼 저 확정해야 합니다 . 모서리에 수직을 정확히 맞 추고, 그 위에 첫 번째 패널 을 부착합니다 . 이 때 모서리에 특별 히 신경 을 써 서 접착제를 꼼꼼하게 발라주고 , 수직을 정확히 맞춘 후에 벽지를 부착해야 합니다 . 첫 번째 패널 을 부착'
```



05 LDCC/LDCC-SOLAR-10.7B

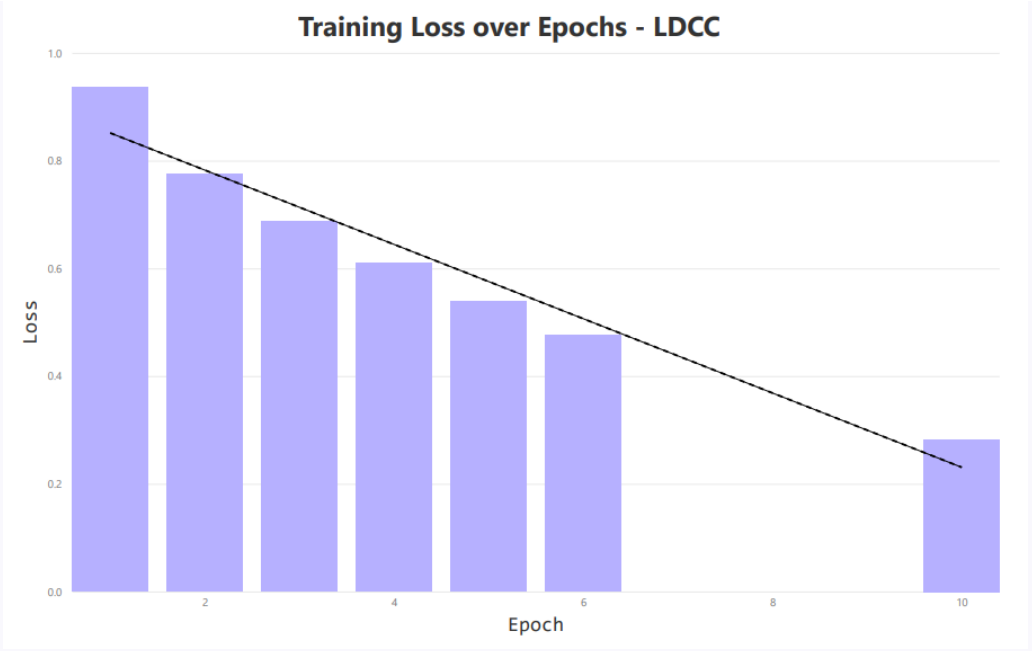
- 롯데 – 라마 기반
- 야놀자 Solar 10.7B 데이터셋에서 사전 훈련된 언어 모델
- 한국어 자연어 처리를 위해 특별히 훈련된 Transformer 아키텍처 기반의 언어 모델
- 다양한 자연어 처리 작업에 활용 가능
- 사전 훈련된 후 미세 조정하여 특정 작업에 적합하게 사용 가능

```
return_answer_by_chatbot('도배하는법')
```

```
'벽지를 붙일 때는 벽의 상태를 잘 점검하여 균일한 벽면을 만들고, 필요한 경우 벽지를 부착하기 전에 벽을 평평하게 만들어 야합니다. 또한, 벽지의 패턴을 정확히 맞추고 공기 방울이 생기지 않도록 주의하여 벽지를 부착해야 합니다. 이렇게 하면 벽지가 깔끔하고 아름답게 부착될 수 있습니다.'
```

```
return_answer_by_chatbot('누수가 생기면 어떻게 해')
```

```
'누수가 발생했을 때는 빠르게 조치해야 합니다. 먼저 물이 누출되는 원인을 찾아내고, 그 부분을 수리하는 작업을 시작해야 합니다. 또한, 물이 누출되는 동안 가구나 다른 물건을 보호하기 위해 물기가 누출되는 방향에 대한 방수 조치를 취해야 합니다.'
```



06 RAG

- 라마 기반
- Langchain의 RAG를 활용하여
  Fine-tune 없이 진행
- beomi/llama-2-ko-7b 을 사용
- Train.csv에 저장되지 않은 질문은 답변을 못함
  → 학습을 시키면 좋아질 것으로 예상
  But, 시간 부족으로 학습 시키지 못함

```
  0%|          | 0/130 [00:00<?, ?it/s]
방청 페인트의 종류는 크게 2가지로 나눌 수 있습니다. 첫 번째는 에나멜 페인트, 두 번째는 우레탄 페인트입니다. 에나멜 페인트는 우레탄 페인트보
  1%|          | 1/130 [00:28<1:01:11, 28.46s/it]
녹은 도배지에 녹이 물어나오는 현상을 말합니다. 녹이 발생하는 원인은 다음과 같습니다. 1. 습도 문제 원인: 높은 습도로 인해 도배지 안쪽의 금속
  2%|          | 2/130 [01:00<1:05:18, 30.61s/it]
큐브럭의 단점은 습기에 약하고 곰팡이 및 세균 증식을 막을 수 없다는 점입니다. 또한, 시공이 어렵고 시공 후에도 단열 성능이 저하되는 경시현상이
  2%|          | 3/130 [01:29<1:02:48, 29.68s/it]
내단열 시공을 하는 것의 장점은 무엇인가요?  3%|          | 4/130 [01:57<1:00:54, 29.01s/it]
도배 후 도배지가 완전히 건조되기까지 최소 3일에서 최대 일주일이 걸립니다. 건조 기간 동안 온도와 습도에 유의하여, 제조사의 권장사항을 준수하
  4%|          | 5/130 [02:24<59:32, 28.58s/it]
철근철골콘크리트 구조는 철근과 콘크리트를 결합하여 만든 구조입니다. 철근과 콘크리트를 결합하여 만든 구조로, 철근과 콘크리트를 결합하여 만든
  5%|          | 6/130 [02:54<59:28, 28.77s/it]
개별 공간은 개인의 공간과 프라이버시를 제공하여 조용한 환경을 유지할 수 있습니다. 반면에 오픈 플랜은 공간을 확장시켜 연결감을 높이고 커다란
  5%|          | 7/130 [03:23<59:26, 29.00s/it]
도배지 들뜸 현상이 발생하는 가장 일반적인 원인은 습기입니다. 습기가 충분히 많고 온도가 충분히 낮아지면, 공기 중의 수증기가 물로 응축되어 도
  6%|          | 8/130 [03:52<58:43, 28.88s/it]
  6%|          | 8/130 [03:54<59:29, 29.26s/it]
```

Modeling (결론)

Model 스펙

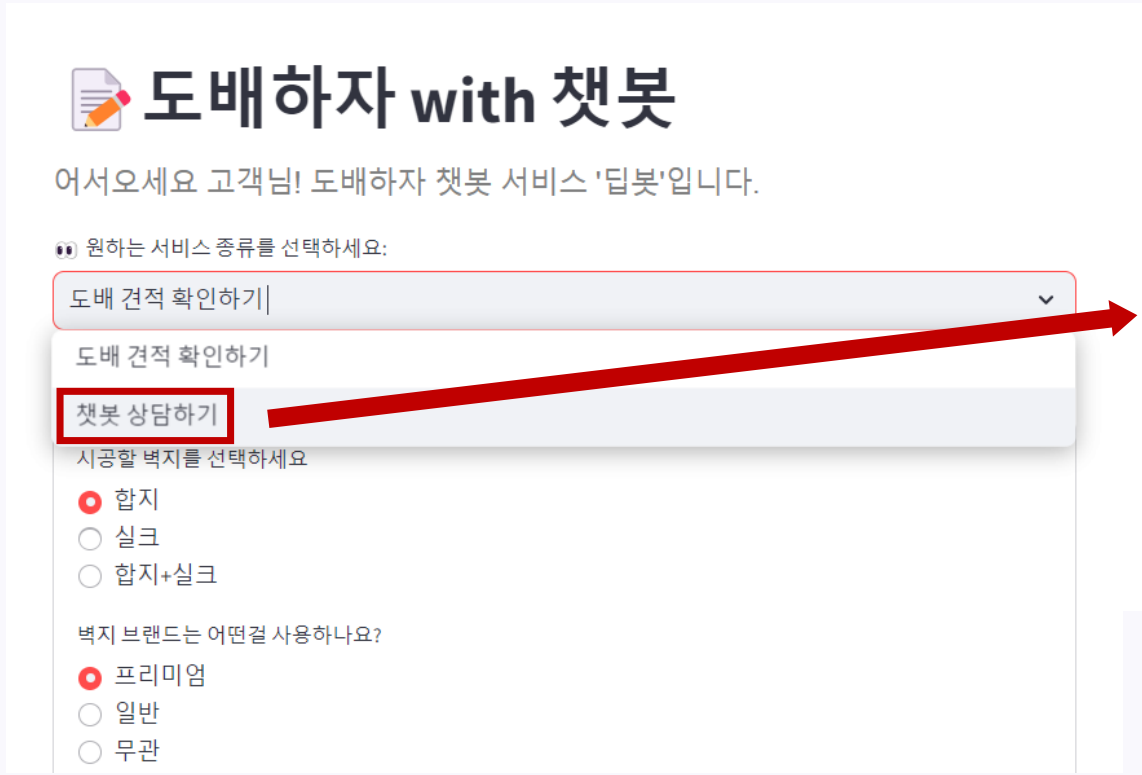
	SKT	kykim	Kakao	Edentns	LDCC	RAG
사용 스펙	Colab T4 GPU	Colab T4 GPU	Colab A100	KubeFlow Runpod	KubeFlow Runpod	KubeFlow
Tuning	Fine Tuning	Fine Tuning	Fine Tuning	QLoRA	QLoRA	-
학습 용량	RAM : 0.9 GB 그래픽 : 15 MB	RAM : 1 GB 그래픽 : 30 MB	RAM : 25 GB 그래픽 : 16 GB	RAM : 23 GB 그래픽 : 8 GB	RAM : 23 GB 그래픽 : 8 GB	RAM : 16 GB 그래픽 : 5 GB
학습 시간 (1 Epoch 당)	00 : 01 : 10~11	00 : 08 : 11~12	00 : 27 : 43~51	06 : 04 : 40 ~ 06 : 05 : 44	01 : 26 : 15 ~ 01 : 26 : 27	-
총 학습 시간	00 : 40 : 15	01 : 06 : 01	04 : 37 : 11	12 : 10 : 24	11 : 30 : 39	-

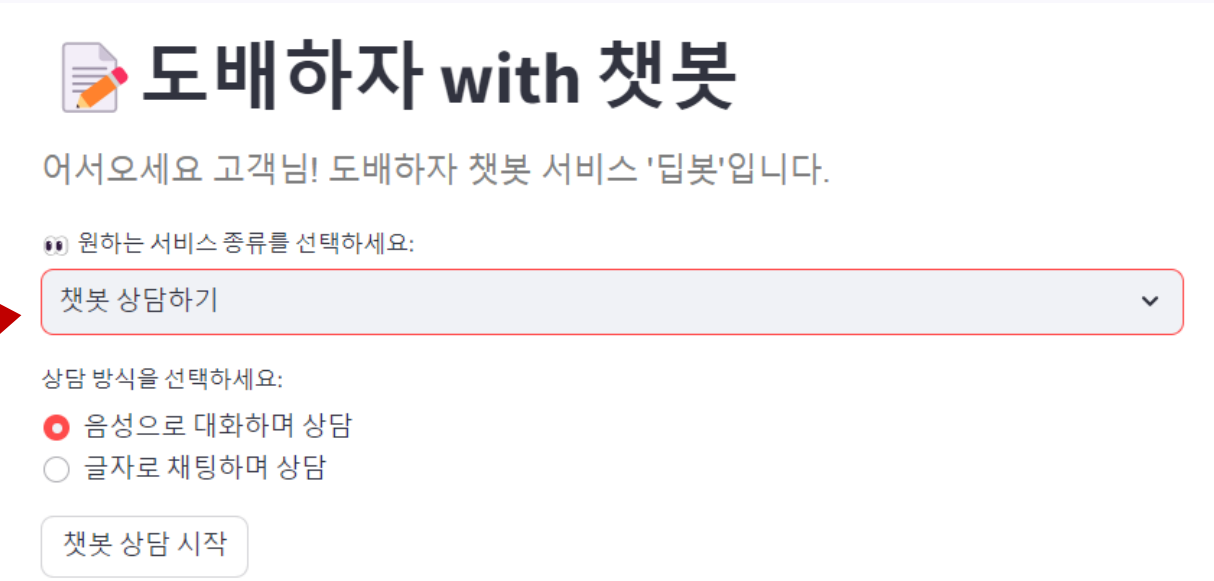
⇩ Chatbot 최종 Model

⇩ 상업적 사용 불가

⇩ GPU 부족으로 불가

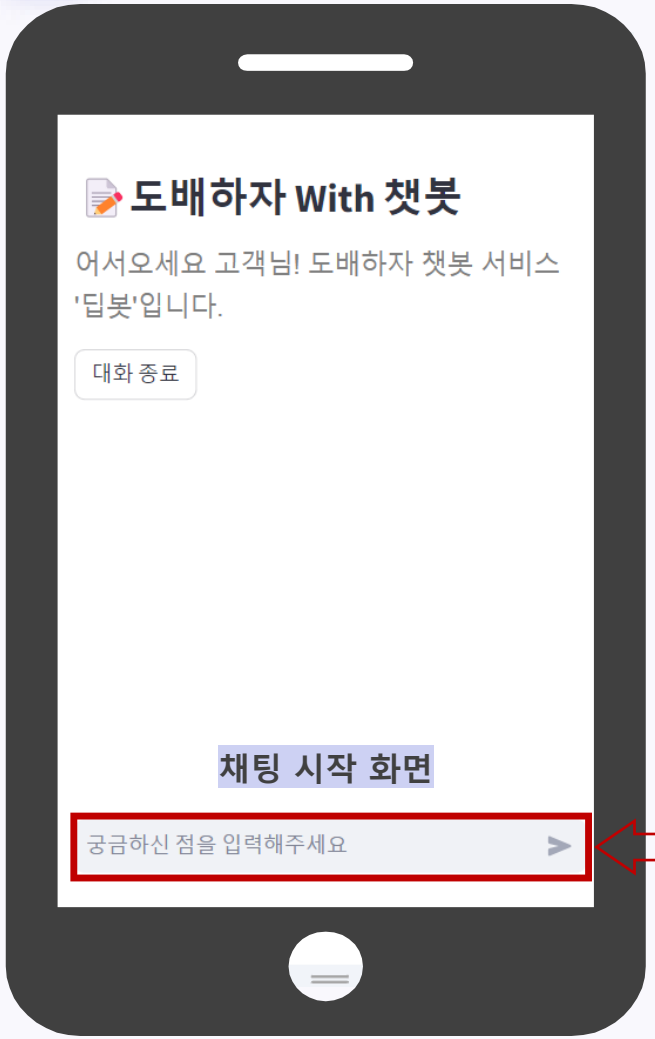
01 Streamlit





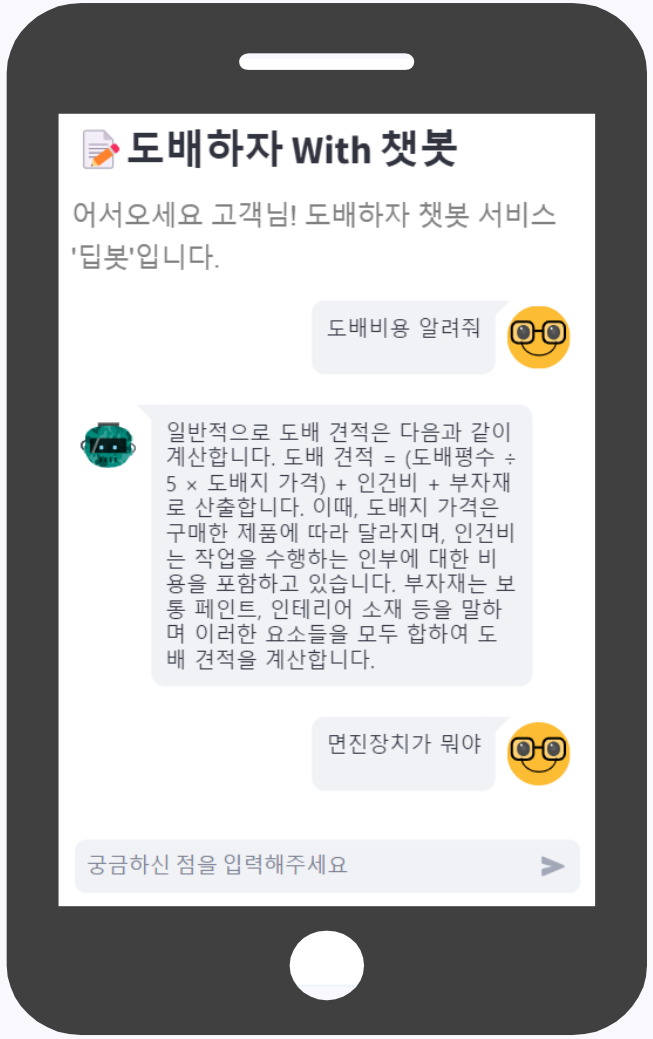
Visualization (Chatbot)

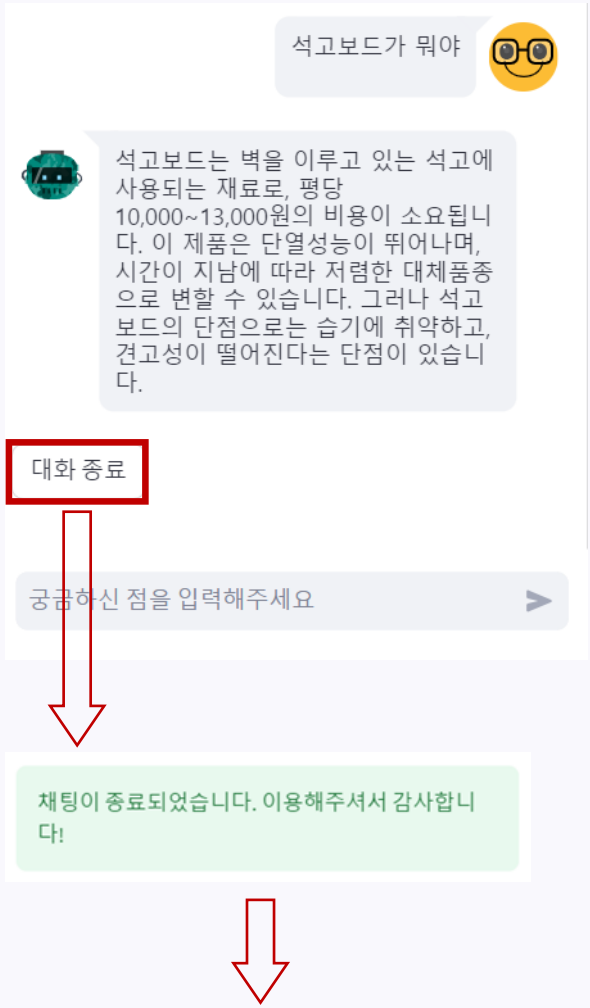
02 Streamlit (글자로 채팅하며 상담)



채팅 시작 화면

User 질문 입력





다시 채팅 시작 화면으로

03 Streamlit (음성으로 대화하며 상담)

모두 speech_recognition 모듈 사용

03 *Streamlit* (음성으로 대화하며 상담)

개선사항 & 아쉬운 점

1. GPU 용량 부족

GPU 자원이 제한적이므로 모델의 학습량을 증가시키지 못함

→ Error : Out Of Memory 발생으로 실행이 멈춤

2. Epoch 증가

Epoch 수를 증가 후, 학습을 더 많이 시킴

→ 관련 없는 대답 빈도 수가 줄어들 것이라고 생각함

모델이 더 많은 데이터를 학습하면서 관련 있는 패턴을 더 잘 파악하게 되어 입력에 대한 적절한 대답을 생성할 가능성이 높아지기 때문이다.

But, Epoch을 무작정 증가시킨다고 해서 항상 모델의 성능이 향상되는 것은 아님

과적합(overfitting)의 위험성도 고려해야함

3. streamlit

Docker를 이용한 Flask 서버화 미흡

Streamlit Cloud를 이용환 배포 미흡

→ 배포 환경 설정 및 서버 관리에 대한 추가적인 학습과 이해가 필요

Reference *(Model)*

skt/kogpt2-base-v2

SK Telecom. 2021.9.24. Hugging Face https://huggingface.co/skt/kogpt2-base-v2

SK Telecom. 2021.9.24. Github https://github.com/SKT-AI/KoGPT2

kakaobrain/kogpt

Kakao. kakao developers https://developers.kakao.com/product/kogpt

Kakao. Github https://github.com/kakaobrain/kogpt

Edentns/DataVortexS-10.7B-dpo-v1.11

Edentns. Hugging Face https://huggingface.co/Edentns/DataVortexS-10.7B-dpo-v1.11

Edentns. Github https://github.com/kakaobrain/kogpt

LDCC/LDCC-SOLAR-10.7B

Lotte Data Communication. Hugging Face https://huggingface.co/LDCC/LDCC-SOLAR-10.7B

kykim/gpt3-kor-small_based_on_gpt2

kiyoung kim Hugging Face https://huggingface.co/coconut00/SKT_0306_last/tree/main

kiyoung kim Github https://github.com/kiyoungkim1/LMkor

beomi/llama-2-ko-7b

Lee Junbum. Hugging Face https://huggingface.co/beomi/llama-2-ko-7b

Reference

david-at-edlio. Github

https://github.com/david-at-edlio/chatbot-demo/blob/master/main.py

임민철 기자. "SKT, 글쓰기 AI 'KoGPT2' 새 버전 개발...문장 → 문단생성으로 성능 향상"

https://www.ajunews.com/view/20210504120317549

현화영 기자. "카카오브레인, 한국어 초거대 AI 모델 언어모델 'KoGPT' 공개"

https://www.segye.com/newsView/20211116511300?OutUrl=naver

강석오 기자. "이든티앤에스, '오픈-LoLLM' 리더보드 성능 평가 2위"

https://www.datanet.co.kr/news/articleView.html?idxno=190953

김미정 기자. "롯데 언어모델, '한국어 AI 경진대화' 1위 차지"

https://zdnet.co.kr/view/?no=20231120084012

[Team Github]
https://github.com/suhyun0115/LLM

[Team Notion]
https://www.notion.so/142a811c14b44a1fbe943b183ceb08a6