

Interview Chatbot

멋쟁이 사자처럼 실전프로젝트2 2팀

김종은 | 천세현 | 김설아 | 장윤정 | 안도형

Contents

01 프로젝트 개요

02 프로젝트 목표

03 프로젝트 진행과정

04 모델 및 데이터셋

05 프로젝트 실험 및 결과

- 1) 프롬프트 구성에 따른 성능 평가/비교
- 2) 모델 학습 방식에 따른 성능 평가/비교

06 프로젝트 시연

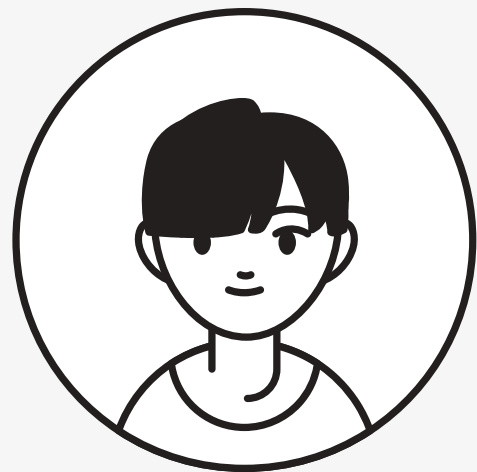
07 스터디

08 프로젝트 발전 방안

02 프로젝트 목표

면접 대비를 위한 면접 챗봇

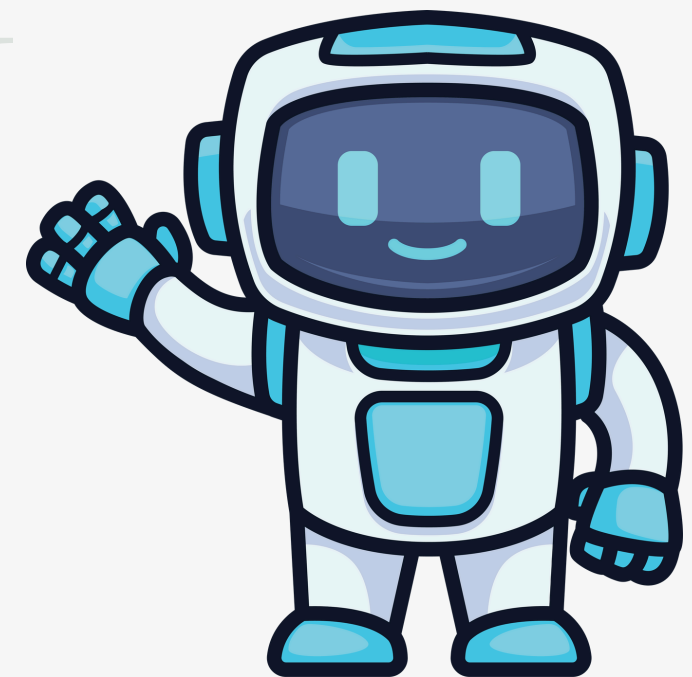
- 면접 준비를 위해서는 예상 질문을 뽑아보고 답변을 준비하는 것이 중요
- 면접 챗봇이 있다면 스스로 질문과 답변을 주고 받으며 준비가 가능



USER

지원동기를 말씀해주세요.

저는 어릴 적부터 컴퓨터 프로그래밍을 꿈꿔왔습니다. 그래서 이 직무에 지원하게 되었습니다. 제가 이 회사에 입사하게 된다면 최선을 다해서 회사에 도움이 되는 직원이 되도록 노력하겠습니다.



03 프로젝트 진행과정

01

모델 및 데이터셋 탐색

- KoGPT2 모델 사용
- AIHUB 채용면접 인터뷰 데이터 활용

02

개발 및 성능 확인

- 다양한 방법으로 파인튜닝 진행
- BLEU, BERTScore, Human score 등으로 평가

03

챗봇 서비스화

- Flask & Docker 기반의 챗봇 서비스화

04 모델 및 데이터셋

KoGPT2 & AIHUB 채용면접 인터뷰 데이터

KoGPT2

모델 특징 :

- SK텔레콤이 오픈한 GPT-2 기반의 한국어 언어 모델로, 한국어 데이터셋으로 학습되어 한국어 자연어 처리에 강점
- 한글의 고유한 문법 구조(조사, 어미 등)를 이해하고 처리하는 데 최적화

모델 선정 이유 :

- 한국어 데이터셋을 사용할 예정이었기 때문에 한국어 처리 능력이 뛰어난 모델이 필요
- 기본적으로 pre-training 된 모델을 사용해야 어느정도 성능이 보장될 거라 생각

AIHUB 채용면접 인터뷰 데이터

데이터셋 특징 :

- 실제 면접 기반 데이터로 질문과 답변으로 구성되어있고 다양한 직군, 성별, 상황에 맞게 세분화 되어있음

데이터셋 선정 이유:

- 한국어 데이터셋
- 데이터의 구조가 질문, 답변으로 명확하게 구분되어 있음

05. 프로젝트 실험 및 결과

1) 프롬프트 구성에 따른 성능 평가/비교

1) 프롬프트 구성에 따른 성능 평가/비교

다양한 형식의 프롬프트 구성하여 학습 진행

프롬프트

- 1 질문+ 답변
- 2 질문 + 요약 답변
- 3 요약 답변 + 답변
- 4 경력 + 질문 + 답변
- 5 직업 + 경력 + 질문 + 답변
- 6 직업 + 경력 + 질문 + 요약 답변
- 7 질문 + 경력 + 직업 + 답변

+

모델 파라미터

Learning rate : 5e-5

Batch Size : 32

Epoch : 6

Trainer 활용 (1-6번)

1) 프롬프트 구성에 따른 성능 평가/비교

모델별 질문-응답 답변 예시

정보통신 분야에 지원한 동기가 무엇인가요?

질문 + 답변 모델

제가 이 분야에 관심을 갖게 된 것은 대학교 다닐 때 전공이 전산통계학이었습니다. 그래서 저는 이 전공을 선택하게 되었습니다. 제가 전산 통계학을 전공하게 된 계기는 대학 졸업 후에 바로 취직이 가능했기 때문입니다. 제가 취직을 하게 된다면 저의 커리어를 쌓을 수 있는 좋은 기회가 될 것이라고 생각합니다. 그리고 제가 앞으로도 계속 이 일을 하고 싶기 때문에 이 직종에 지원하게 됐습니다.

질문 + 요약 답변 모델

저는 어릴 적부터 컴퓨터 프로그래밍을 꿈꿔왔습니다. 그래서 이 직무에 지원하게 되었습니다. 제가 이 회사에 입사하게 된다면 최선을 다해서 회사에 도움이 되는 직원이 되도록 노력하겠습니다.

1) 프롬프트 구성에 따른 성능 평가/비교

성능 평가지표

BERT Score

텍스트 유사도를 평가하기 위해 BERT 모델의 임베딩 활용 / 단어 간 의미적 유사성 비교, 문장 구조와 문맥 고려 가능

BLEU

기계 번역 성능 평가를 위해 고안된 지표로 n-gram의 일치를 기반으로 평가 / 생성된 텍스트가 참조 텍스트와 얼마나 유사한지 점수 측정

Rouge-1

생성된 텍스트와 참조 텍스트 간의 단어 단위 일치 계산 / 단순히 개별 단어의 정확도 측정, 단일 단어 기반 (n=1)

Rouge-2

단어 쌍 (bigram)의 정확도 계산 / 단어의 연결성을 평가해 보다 정교한 평가 가능

Rouge-L

생성된 텍스트와 참조 텍스트 간의 Longest Common Subsequence (LCS)를 기반으로 평가 / 문장 구조와 순서를 유지한 채 일치하는 부분을 측정

1) 프롬프트 구성에 따른 성능 평가/비교

BLEU, BERT, ROUGE 스코어

	Bert Score	Rouge-1	Rouge-2	Rouge-L	BLEU
요약 + 답변	72.4029	5.598141	0.133037	5.598141	6.8772
질문 + 답변	68.3263	3.908684	0.110336	3.905569	2.8024
경력 + 질문 + 답변	72.9499	5.65645	0.141448	5.65022	8.1975
직업 + 경력 + 질문 + 답변	69.7682	4.47448	0.114161	4.47448	3.4636
질문 + 경력 + 직업 + 답변	73.90	30.6589	9.1081	24.2107	1.7672
질문 + 요약	62.9716	1.128836	0.008356	1.128836	0.129
직업 + 경력 + 질문 + 요약	65.067	1.182884	0.009527	1.182884	0.0395

1) 프롬프트 구성에 따른 성능 평가/비교

BLEU, BERT, ROUGE 스코어 비교 결과

평가 결과 비교

- '정답'에 대한 결과 > '요약'에 대한 결과
- 더 정확한 응답이 더 높은 평가 지표 기록

텍스트 길이와 성능

- 텍스트가 많을수록 대부분의 지표가 상승
- 패딩토큰이 적고 무시되는 토큰이 적을수록 성능 향상

정보량과 성능 관계

- '경력+질문+답변' > '질문+답변'
- 비교할 대상이 더 많을수록 평가 점수 상승

BLEU, METEOR, ROUGE는 모두 통계적 원리 기반 점수인 만큼
통계적 평가의 한계로 인한 HUMAN SCORE의 필요성 대두

1) 프롬프트 구성에 따른 성능 평가/비교

Human Score

팀원들이 직접 스코어를 매겨 생성된 응답 평가 진행하였으며, 최종 점수는 아래와 같이 평균값을 계산하였음

$$\text{Final Score} = \frac{\sum_{i=1}^n D_i + \sum_{i=1}^n Q_i + \sum_{i=1}^n C_i}{n}$$

where:

D_i = Document Accuracy score given by the i -th evaluator (0-2),
 Q_i = Question Relevance score given by the i -th evaluator (0-2),
 C_i = Context Consistency and Fluency score given by the i -th evaluator (0-2),
 n = Number of team members (evaluators).

배점	문장 정확도	질문 관련성	문맥 일관성 및 답변 풍부함 여부
0	문장이 말이 안됨	질문과 전혀 관련이 없거나 완전 무의미	답변이 문맥과 전혀 연관성이 없으며, 매우 단편적이거나 부실함
1	문장 구조가 이상한 부분이 있음	질문과 약관 연관은 있으나 질문과 무관하거나 부적절한 내용 포함	답변이 문맥과 어느 정도 연관성이 있으며, 정보의 양이 기본 수준임
2	문장이 말이 됨	질문의 핵심에 완벽히 부합하여 모든 요소를 정확하게 다룸	답변이 문맥과 완전히 연관성을 유지하며, 내용이 풍부하고 구체적임

1) 프롬프트 구성에 따른 성능 평가/비교

Human Score를 활용한 성능 비교 결과

Question 1. 정보통신 분야에 지원한 동기가 무엇인가요?

Best : Worst :

배점	질문+답변	질문 + 요약	요약 + 답변	경력 + 질문 + 답변	직업 + 경력 + 질문 + 답변	직업 + 경력 + 질문 + 요약	질문 + 경력 + 직업 + 답변
문장 정확도	10	10	7	9	8	7	4
질문 관련성	6	8	7	9	6	4	3
문맥 일관성 및 답변 풍부함 여부	4	5	6	7	6	4	4
Final Score (AVG)	4	4.6	4	5	4	3	2.2

1) 프롬프트 구성에 따른 성능 평가/비교

실제 답변 예시

● Best : 경력 + 질문 + 답변

Generated Answer :

제가 이 분야에 관심을 가지게 된 이유는 제가 정보통신에 관심이 많기 때문입니다. 저는 어렸을 때부터 컴퓨터를 많이 다루어 왔기 때문에 컴퓨터를 잘 다룰 줄 압니다. 그래서 이 회사에 지원하게 되었습니다. 그리고 제가 가장 잘 할 수 있는 일이 이 정보통신 분야라고 생각하기 때문에 지원을 하게 되었습니다.

● Worst : 질문 + 경력 + 직업 + 답변

Generated Answer :

저는 이 회사에 입사를 하게 된다면 가장 가고 싶은 부서는 개발 부서입니다. 개발 부서에 들어가고 싶다는 생각이 들었기 때문에 가장 일하고 싶은 부서라고 생각합니다. 개발부서는 제가 하고 싶은 일을 할 수 있는 곳이기 때문에 굉장히 좋은 부서가 될 수 있다고 생각하기 때문입니다.

2) 모델 학습 방식에 따른 성능 평가/비교

성능 평가 결과

	Bert Score	Rouge-1	Rouge-2	Rouge-L	BLEU
ALL Layers	0.6177	0.1216	0.0279	0.1171	0.0080
Last Layer	0.5928	0.0805	0.0197	0.0780	0.0034

ALL Layer

저는 대학 시절 팀 프로젝트를 진행한 경험이 있습니다. 당시 저희 과는 디자인과 학생들끼리 팀을 이루어서 진행했습니다. 저희는 디자인 전공 학생들과 함께 팀을 이뤄서 프로젝트에 참여했었는데요. 저희가 디자인 전공을 하고 있는 친구들은 각자 전공에 대한 지식이 부족했었습니다. 그래서 저희의 디자인 팀원들은 서로 다른 전공의 친구들끼리는 서로에 대해 잘 알지 못했습니다. (중략) 이러한 경험을 바탕으로 앞으로도 제가 디자인 업무를 할 때 서로 부족한 부분을 채워줄 수 있는 그런 인재가 될 수 있도록 노력하겠습니다. 감사합니다.

Last Layer

어 저는 어 제가 어 어 그 어 갈등 상황이 어 해결이 된다면 어 다른 사람들과 어 갈등이 생겼을 때 어 서로 어 대화를 통해서 어 갈등을 해결했던 경험이 있습니다. 어 그래서 어 이 갈등 상황에서는 어 여러 가지 어 문제들이 어 발생을 하는데 어 일단 어 먼저 어 갈등의 원인을 어 파악을 하고 어 해결을 하는 것이 어 가장 중요하다고 생각을 합니다. (중략) 뭘 어떻게 해결해야 될까 어 이런 고민이 많이 있습니다.

1) 프롬프트 구성에 따른 성능 평가/비교

Human Score를 활용한 성능 비교 결과

Question 2. 갈등 상황을 해결한 경험을 이야기해주세요

Best : Worst :

배점	질문+답변	질문 + 요약	요약 + 답변	경력 + 질문 + 답변	직업 + 경력 + 질문 + 답변	직업 + 경력 + 질문 + 요약	질문 + 경력 + 직업 + 답변
문장 정확도	10	8	10	10	4	9	4
질문 관련성	10	3	10	7	4	3	8
문맥 일관성 및 답변 풍부함 여부	10	3	7	8	5	4	6
Final Score (AVG)	6	2.8	5.4	5	2.6	3.2	3.6

1) 프롬프트 구성에 따른 성능 평가/비교

실제 답변 예시

● Best : 질문 + 답변

Generated Answer :

저는 대학 시절 팀 프로젝트를 진행한 경험이 있습니다. 당시 저희 과의 특성상 조별 과제를 해야 하는 경우가 많았는데요. 이럴 경우 팀원들 간의 불화가 가장 큰 문제였습니다. 서로 다른 의견을 가지고 있기 때문에 의견 조율이 가장 어려웠습니다. 하지만 이 문제를 해결하기 위해선 서로의 의견을 경청하는 것이 가장 중요하다고 생각했습니다. (중략) 또한 이러한 시간을 통해 서로에 대해 알아갈 수 있었고 서로에게 대해 알아가게 된다면 서로가 서로를 더 신뢰하고 더 좋은 방향으로 나아갈 수 있을 거라 생각합니다.

● Worst : 질문 + 경력 + 직업 + 답변

Generated Answer :

네 저는 어떤 일을 할 때 갈등이 생길 수 있다고 봅니다. 그럴 때는 그 갈등의 원인이 무엇인지를 먼저 잘 파악해야 된다고 생각합니다. 그래서 어떤 갈등이 생겼을 때 그 원인을 잘 파악하고 그 갈등을 해소할 수 있는 방법을 잘 강구해야 된다고 생각합니다. 그렇게 함으로써 어떤 갈등이나 불협화음이 생기지 않을 수 있기 때문에 그렇습니다. 항상 갈등은 피할 수 없는 것 같습니다. 그래서 갈등이 생긴다면 그것을 잘 해소하는 것이 가장 좋은 방법이라 생각되기 때문에 어떤 갈등은 어떤 방법으로 풀어나가야 된다고 보고 있습니다. (중략) 그렇게 할 때는 항상 어떤 방법을 동원해서라도 갈등을 해결해야 된다는 그런 생각을 항상 가지고 있어야만 그 갈등은 해소될 수

05. 프로젝트 실험 및 결과

2) 모델 학습 방식에 따른 성능 평가 및 비교

2) 모델 학습 방식에 따른 성능 평가/비교

All Fine Tuning VS Freezing

Model 1

- Fine-Tuning All Layers
- 기존 사전학습된 가중치에서 벗어나 새로운 데이터 셋 특성을 잘 반영할 수 있음
→ 데이터가 기존 사전학습된 데이터와 크게 다를 경우, 유리함
- 높은 자원이 요구되며, 사전학습된 지식 손실 위험이 존재함

Model 2

- Fine-Tuning Last Layer Only
- 사전학습된 지식을 유지할 수 있으며 파라미터의 수가 줄어 메모리 사용량과 계산 비용을 아낄 수 있음
→ 데이터셋이 작을 경우, 사전학습된 모델의 이점을 극대화
- 새로운 데이터셋의 특성을 깊이 반영하지 못 할 가능성이 존재함

```
GPT2LMHeadModel(  
  (transformer): GPT2Model(  
    (wte): Embedding(51200, 768)  
    (wpe): Embedding(1024, 768)  
    (drop): Dropout(p=0.1, inplace=False)  
    (h): ModuleList(  
      (0-11): 12 x GPT2Block(  
        (ln_1): LayerNorm((768,), eps=1e-05, elementwise_affine=True)  
        (attn): GPT2SdpaAttention(  
          (c_attn): Conv1D(nf=2304, nx=768)  
          (c_proj): Conv1D(nf=768, nx=768)  
          (attn_dropout): Dropout(p=0.1, inplace=False)  
          (resid_dropout): Dropout(p=0.1, inplace=False)  
        )  
        (ln_2): LayerNorm((768,), eps=1e-05, elementwise_affine=True)  
        (mlp): GPT2MLP(  
          (c_fc): Conv1D(nf=3072, nx=768)  
          (c_proj): Conv1D(nf=768, nx=3072)  
          (act): NewGELUActivation()  
          (dropout): Dropout(p=0.1, inplace=False)  
        )  
      )  
    )  
    (ln_f): LayerNorm((768,), eps=1e-05, elementwise_affine=True)  
  )  
  (lm_head): Linear(in_features=768, out_features=51200, bias=False)
```

06 시연

Hi, I'm InterviewBuddy! 😊

InterviewBuddy는 채용 면접 연습을 도와주는 AI 챗봇이에요
예상 면접 질문을 입력하면, InterviewBuddy가 답변을 제공해요
면접 준비를 하고 있는 당신, 바로 사용해 보세요!

📖 사용 방법

- 1 예상 질문을 입력하세요
- 2 "Send" 버튼을 눌러 AI 답변을 확인하세요
- 3 새로운 대화를 시작하려면 "New Chat"을 클릭하세요

Start Chatting

InterviewBuddy 😊

New Chat

본인의 강점이 무엇이라고 생각하시나요

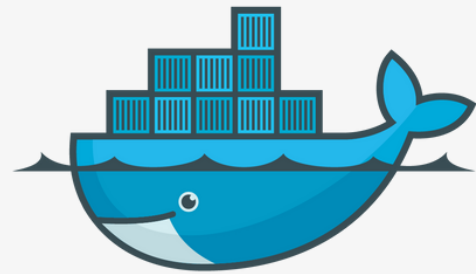
저는 성격이 내성적이지만 사람 만나는 걸 좋아하고 바깥 활동을 좋아하는 편이라 여러 사람들과 같이 어울려 사이클링 등산 등 사 모임에 적극적으로 활동하고 성격이 모나지 않아 사람들과 잘 어울린다는 소리를 많이 들었습니다. 이런 성격이 장점인 것 같습니다. 단점으로는 주어진 업무 수행의 과도한 집중력으로 인해서 조금해 하거나 서두르는 경향이 있다고 합니다. 그래서 주어진 과제와 해결해야 될 과제가 있을 때 그 과제에 대해서 혼자 해결하려는 개인 성향이 강하다고 합니다.

Enter your question...

Send 💬

07 스터디

Docker란?



docker

Write Once, Run Anywhere !

컨테이너 기반 가상화 기술

애플리케이션과 의존성을 묶어 컨테이너로 배포

Docker의 장점

이식성

어떤 환경에서도 동일하게 실행 가능

격리성

하나의 컨테이너는 독립적으로 실행

효율성

가상 머신보다 가벼운 환경, 적은 리소스 사용

빠른 배포

컨테이너 시작, 중지 명령어로 빠른 배포, 스케일링 가능

활용 방안

협업 시 개발 환경 통합

데이터 전처리 자동화

NLP 모델 개발 및 서빙

배포 자동화

07 스터디

Flask로 구현한 챗봇을 Docker Image로 만들어 배포하기

Dockerfile

```
FROM python:3.12-slim as builder

WORKDIR /app

COPY requirements.txt /app/requirements.txt
RUN apt-get update && apt-get install -y \
    build-essential gcc \
    && pip install --no-cache-dir -r requirements.txt \
    && apt-get clean \
    && rm -rf /var/lib/apt/lists/*

FROM python:3.12-slim

WORKDIR /app

COPY --from=builder /usr/local/lib/python3.12 /usr/local/lib/python3.12

COPY . /app

ENV FLASK_APP=app.py
ENV FLASK_RUN_HOST=0.0.0.0
ENV FLASK_ENV=production

VOLUME /app/model

CMD ["python3", "-m", "flask", "run", "--host=0.0.0.0"]
```

Image Build

\$ docker build -t [이미지이름]

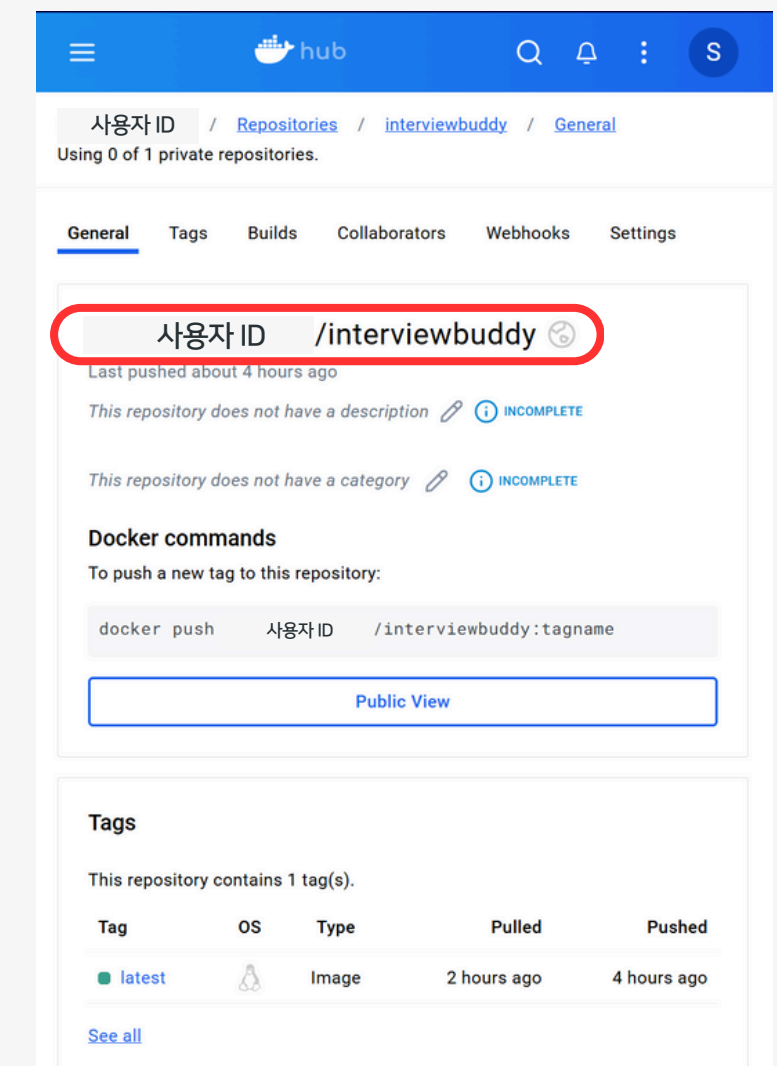
Image Tagging

\$ docker tag [이미지ID] [사용자ID]/[이미지이름]:tag

Image Push

\$ docker push [사용자ID]/[이미지이름]:tag

Docker Hub



08 프로젝트 발전방향

프롬프트 엔지니어링

- 프롬프트에 조건이나 제한 사항을 추가해서 조금 더 자세하고 주제에서 벗어나지 않는 대답을 생성하도록 엔지니어링 해보면 좋을것이라 생각
- hallucination 방지

데이터 전처리

- 구어체 정리 등 음성 데이터에서 추출된 데이터인 만큼 학습 시키고 읽기에 더 편하게 데이터를 정리한다면 더 좋은 학습 결과와 답변이 생성될 것으로 기대

파라미터 튜닝

- 배치 사이즈의 제한도 많았고 조금 더 다양한 파라미터를 실험해보면 더 나은 결과를 찾을 수 있을 것으로 기대



질문 및 답변

궁금한 사항을 질문해 주세요.

Thank you

감사합니다.