한솔데코 시즌2 AI 경진대회

도배 하자 QA 솔루션 개발

도배하자와 관련된 깊이 있는 질의응답 처리 능력을 갖춘 AI 모델 개발

- 01. 대회 개요 및 대회 목적
- 02. 데이터 활용
- 03. 평가 전략
- 04. LLM 모델
- 05. RAG
- 06. 결론
- + Appendix

개요

대회 배경 및 목표



도배 하자 및 인테리어 관련 다양한 질문에 응답을 제공하는 AI 모델 개발 고객의 문의에 신속하고 정확하게 답변할 수 있는 시스템을 구축

개요

데이터셋 설명

train.csv (644행)

컬럼명	예시	
질문_1	조적식 구조는 어떤 방식이 있어?	
질문_2	조적식 구조의 다양한 방식에 대해 알려주세요.	
category	건축구조	
답변_1	조적식 구조는 벽돌구조, 블록구조, 석구조가 있습니다.	
답변_2		
답변_3		
답변_4		
답변_5	조적식 구조는 주로 벽돌, 블록, 그리고 석재 같은 자연 원료로 건물을 건설하는 방식을 말합니다. 이는 각각의 자연원료를 사용하여 벽돌 구조, 블록 구조, 그리고 석 재 구조로 나뉘게 됩니다.	

test.csv (130행)

컬럼명	설명
질문	응답을 얻고자 하는 테스트 질문 (Al 모델의 인풋)

submission.csv (130행)

컬럼명	설명
vec_0	
	test.csv 질문에 맞는 응답을 512차원의 벡터로 변환하여 제출
vec_511	

개요

개발 전략 개요

"당장 서비스 가능한" AI 응답 봇 개발





▶ 당장의 리더보드 점수 보다 실제 응답의 퀄리티 & 답변 속도 개선에 포커스

〈 Data 〉

- 데이터 전처리 전수 조사
- 데이터 증강

Model > ______

- · 문제 해결 접근의 모델 선택
- SFT & DPO &

Prompt engineering

⟨RAG⟩ _

- 키워드 요약 문서 생성
- 외부 데이터 수집

01. 대회 개요 및 대회 목적

02. 데이터 활용

03. 평가 전략

04. LLM 모델

05. RAG

06. 결론

+ Appendix

01. 데이터 전처리

전처리 1. 데이터 품질 향상을 위한 데이터 클렌징 진행

수정 전 질문	수정 후 질문	
~도배지를 철저히 건조시키고 교체하는 것이 좋습니다. preventive measures are essential.	~ 도배지를 철저히 건조시키고 교체하는 것이 좋습니다.	
합지벽지의 얼룩을 지우는 방법은 식빵에 <mark>치약기</mark> 나 <mark>주바세제응</mark> 조금 <mark>뭍힌</mark> 후 ~	합지벽지의 얼룩을 지우는 방법은 식빵에 <mark>치약</mark> 이나 <mark>주방세제</mark> 를 조금 <mark>묻힌</mark> 후 ~	
전벽돌은 주로 검은색으로 가공된 벽돌로, 특히 검은색 흙이 적재된 공간에서 매우 높은 온도에서 <mark>굽음</mark> 으로써 제작됩니다.	전벽돌은 주로 검은색으로 가공된 벽돌로, 특히 검은색 흙이 적재된 공간에서 매우 높은 온도에서 <mark>구음</mark> 으로써 제작됩니다.	

〈데이터 품질 문제 발견〉

- 답변과 관련 없는 영어표현
- 맞춤법
- 오타
- 내용 오류 수정



데이터 품질 개선으로 인한 SFT 학습 성능 향상 기대

01. 데이터 전처리

전처리 2. 보다 정확한 LLM 학습을 위해 **질문-답변 일치성** 전수조사 진행

질문	답변	질문–답변 일치성
철골 구조의 <mark>장점</mark> 을 알려줘	철골 구조는 화재 발생 시 철골이 녹아 … <mark>단점</mark> 이 있습니다.	X
누수에 의해 도배지에 <mark>녹이 묻어나</mark> 올 수 있어?	누수에 의해 도배지가 젖 있는 상 태가 지속되면 <mark>곰팡이가 발생</mark> 할 수 있습니다.	X DELETE
벽면에 손상이 있을 경우 도배지 꼬임이 발생할 수 있어?	벽면에 손상이 있을 경우 ~ 꼬임 이 발생할 가능성이 있습니다.	OEL
높은 습도로 인해 몰딩수정이 발 생하는데 <mark>대처할 방법</mark> 이 있을까요?	고습도 환경에서 몰딩 소재가 팽창 하거나 수축하면서 변형될 수 있습 니다.	X DELETE
높은 습도에 의해 석고수정이 발 생할 수 있어?	석고는 습기를 흡수하고 물에 녹 는 성질이 있기에 습도가 높으면 변형될 수 있습니다.	U
		•••

- 〈질문-답변 일치성 전수조사 결과〉

• 질문-답변 총 데이터 : 6,440

• 질문-답변이 잘못 매칭된 데이터: 203

• 삭제 후 데이터 : 6,237

02. 데이터 증강

데이터의 다양한 활용을 위해 TRAIN 데이터의 변형 및 외부 데이터 수집 진행

 키워드
 Train Contents

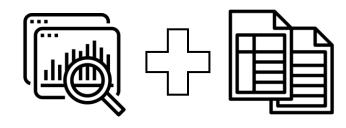
 강화마루
 강화마루 특징

 강화마루
 강화마루 장점

 강화마루 단점
 강화마루 단점

〈Train키워드 기준 답변 요약〉

〈외부 데이터 추가 수집〉



TRAIN 데이터

외부 데이터

- 01. 대회 개요 및 대회 목적
- 02. 데이터 활용
- 03. 평가 전략
- 04. LLM 모델
- 05. RAG
- 06. 결론
- + Appendix

01.성능평가 전략

AI 도배하자 응답 모델을 실제 서비스에 적용할 수준의 성능인지 평가하기 위해 다양한 지표를 추가적으로 사용함

답변 품질

- 1. Human Evaluation 아래 문항에 대해 10점 평가
- (문장 완성도)
 - 다중 질문에 대한 대답을 하였는가
 - 한국어로만 대답하였는가
 - 문장을 반복하지 않는가
 - 문장이 완전한가 (끊김, 비문, 오탈자)
- (내용 정확도)
 - 핵심 키워드가 포함되어 있는가
 - 틀린답을 포함하고 있지 않은가(할루시네이션)
 - 질문과 연관된 RAG 문서를 추출하였는가
 - RAG 문서와 무관한 답변을 하지 않는가

2. Cosine 유사도

• Test 데이터의 30%를 랜덤 샘플하여, 모범답안을 작성하고 내용적 유사도를 평가함

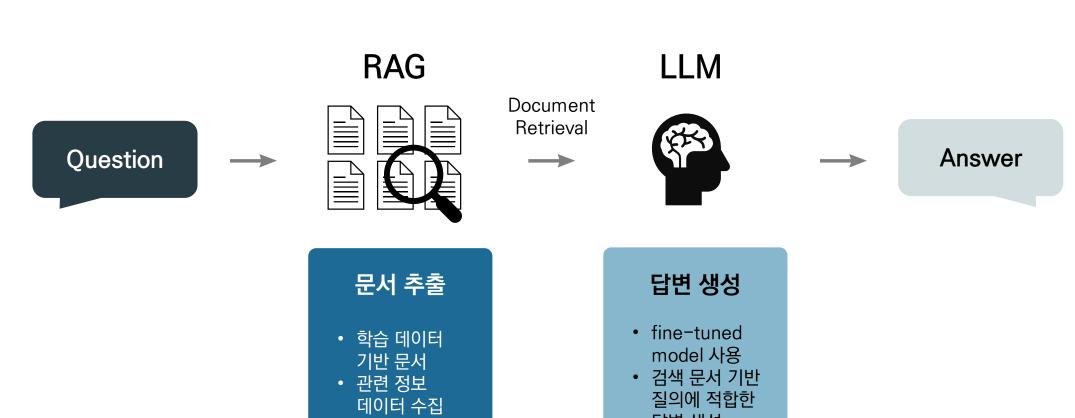
속도

- 1. Throughput
- Test Data 130개의 inference 속도 평가

- 01. 대회 개요 및 대회 목적
- 02. 데이터 활용
- 03. 평가 전략
- 04. LLM 모델
- 05. RAG
- 06. 결론
- + Appendix

01. 서비스 모델 아키텍처

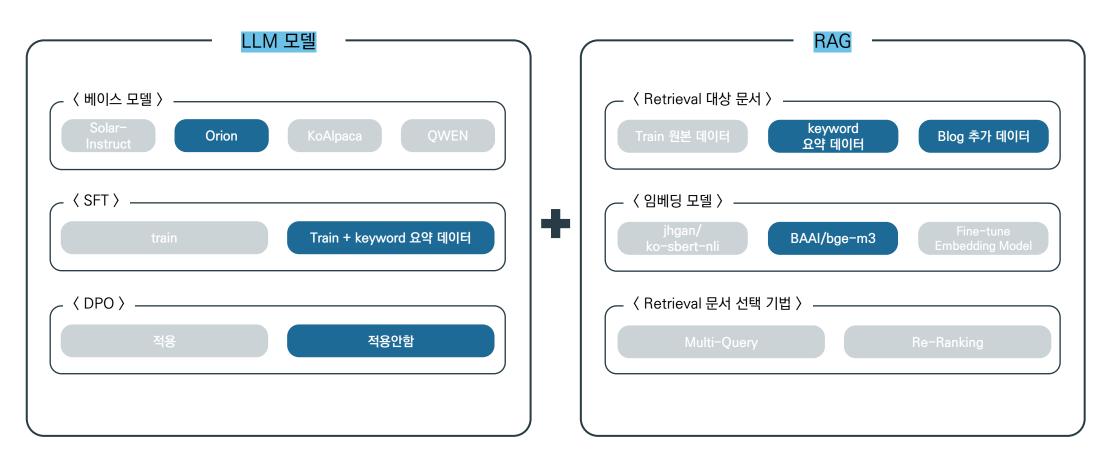
모델 아키텍처



답변 생성

02. 최종 모델 및 적용 기법

자체 성능 평가 기준과 리더보드 점수를 고려해 실험 모델 결과를 비교한 후, 내용의 정확도와 어색함 없는 문장 생성을 중심으로 몇 번의 재현에도 가장 정확하고 안정적인 답변을 보이는 모델과 기법 조합을 최종 선택함



03. SFT (Supervised Fine-tuning)

- 한국어 및 다국어 언어로 사전학습 진행된 거대언어모델 4개 선정
- 거대언어모델의 효과적인 미세조정을 위해 PEFT(Parameter Efficient Fine-tuning) 기법인 LoRA와 양자화 적용
- LoRA(Low-Rank Adaptation): 대부분의 매개변수 가중치는 유지하고 일부만 미세조정하는 fine-tuning 기법
- 양자화(Quantization): 언어 모델의 매개변수를 실수형에서 정수형으로 바꾸어 비트 수를 줄여 모델 사이즈를 줄이는 기법

LLM 모델 탐색

KoAlpaca

- Polyglot-ko 5.8B 기반 한국어 데이터셋으로 학습시킨 모델
- 네이버 지식인을 크롤링 데이터 기반으로 ChatGPT를 통해 새로 생성

OWEN

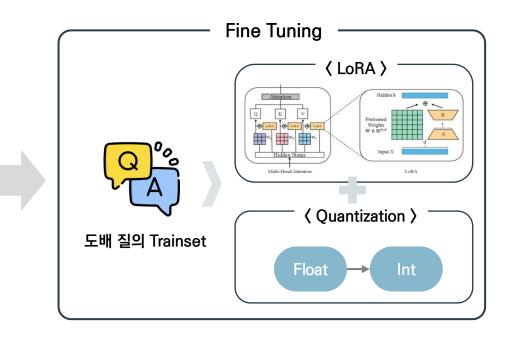
- Alibaba Cloud에서 개발한 다국어 학습시킨 대형 언어모델
- 온라인텍스트, 전문서적, 코드 등 다양한 데이터 수집하여 훈련

Solar

- Upstage가 LLaMA-2 기반으로 파인튜닝한 모델
- Depth-Up-Scaling 방법론을 활용하여 LLM 리더보드에서 1등을 달성

Orion

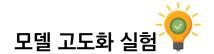
- OrionStarAl가 다양한 분야에서 수집한 다국어 데이터를 훈련시킨 모델
- 종합 평가 및 한국어 테스트에서 뛰어난 성능 발휘

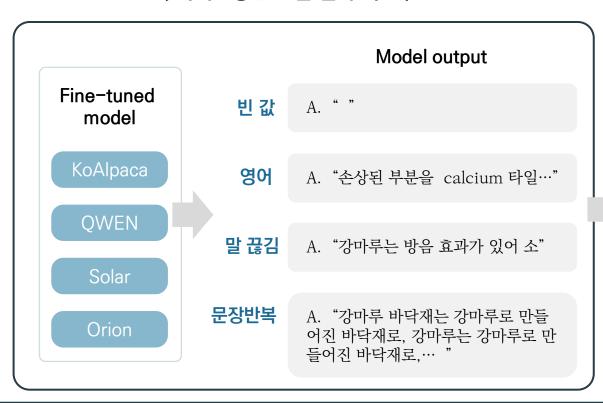


03. SFT - 학습 결과

- 4개 모델에 대해 사전학습 진행 후, 정성 평가로 고도화 포인트 탐색
- 추론 답변 확인 결과, 빈 값 및 영어 답변과 그럴듯한 답변(Hallucination), 말 끊김, 문장반복 출력 확인
- 답변 품질 향상을 위해 DPO 기법, 프롬프트 엔지니어링, 하이퍼파라미터 튜닝 실험 진행

〈 미세조정된 모델 결과 확인 〉





DPO (Direct Preference Optimization)

- 영어 답변이 아닌 한국어 답변만 출력하도록 학습 데이터를 영어로 번역 하여 한글을 chosen으로 영어를 rejected로 구성하여 학습
- 정성 평가 기준(반복없음, 핵심키워드 포함 등)에 부합하는 답변 출력하도 록 평가 점수 기반으로 5점 미만 rejected, 5점 이상 chosen으로 학습

프롬프트 엔지니어링

• "한국어로 답변해주세요", "3문장 내외로 요약해주세요", "모든 질문에 대답해주세요" 등으로 실험

하이퍼 파라미터 튜닝

• Max_length@epoch, temperature, top_k 등 조정하여 성능 실험

03. SFT - 모델 최종 선택

- 4가지 모델에 대해 정성평가를 실시하여 모델 별 특징을 탐색하고 답변 퀄리티가 낮은 모델을 후보에서 제거 (KoAlpaca, QWEN)
- Solar 모델과 Orion 모델의 답변 퀄리티가 유사했으나 모델 배포 후 실제 서비스 적용을 고려하여 추론속도가 빠른 Orion 모델 선택
- Orion-14B-Chat 모델에서 영어 답변과 빈스트링이 출력되어 RAG 활용 시 더 성능이 좋다고 알려진 Orion-14B-Chat-RAG 모델을 추가로 Fine-tuning 하여 한국어 답변만 출력되고 내용 정확도가 좋은 Orion-14B-Chat-RAG 모델로 최종 선택

	KoAlpaca	QWEN	Solar	Ori	ion
최종 모델	KoAlpaca-Polyglot- 5.8B	Qwen1.5-7B-Chat	SOLAR-10.7B- Instruct-v1.0	Orion-14B- Chat	Orion-14B- Chat-RAG
모델 특징	다중질문 답변 못함단어반복존재	- 영어 답변 출력 - 할루시네이션 존재	- 문장 완성도 높음 - 추론 속도 느림	추론 속도 빠름영어 답변 일부 출력빈 스트링 일부 출력	한국어 답변만 출력추론 속도 빠름
정성평가 점수	4.6점	5점	6.2점	7.75점	7.53점
추론 속도 (130개 testset)	28분	29분	22분 33초	13분 30초	11분 24초
학습 속도	2시간 (1epoch)	2시간 40분 (3 epoch)	37분 4초 (1 epoch)	2시간 20분 (3 epoch)	2시간 40분 (3 epoch)

- 01. 대회 개요 및 대회 목적
- 02. 데이터 활용
- 03. 평가 전략
- 04. LLM 모델
- 05. RAG
- 06. 결론
- + Appendix

01. RAG (Retrieval-Augmented-Generation)

- SFT만 진행된 모델의 경우 거짓된 정보를 통해 답변하는 할루시네이션 문제 발생
- 답변의 정확도 문제를 해결하기 위해 문서를 참고하여 답변하는 RAG 기법을 선택함



0. 라돈을 측정하는 데 가장 적합한 지점은 어디인가요?

A. 라돈은 공동주택의 최저층에서 측정하며, 공동주택의 총 세대수에 따라 측정 지점을 결정합니다. 보통 100세대까지는 3세대를 측정하고, 이후 100세대가 증가할 때마다 1세대씩 추가하여 최대 12세대까지 시료를 채취합니다.



SFT

라돈을 측정하는 데 가장 적합한 지점은 알 수 없습니다. 이는 라돈이 지하나 지하실에서부터 상승하여 거주 공간으로 이동하는 특성을 고려하므로 상황마다 다르기 때문입니다.



SFT+RAG

라돈은 공동주택의 최저층에서 측정하며, 공동주택의 총 세대수에 따라 측정 지점을 결정합니다.

100세대일 때는 3세대를 측정하고, 100세대가 증가할 때마다 1세대씩 추가하여 최대 12세대까지 시료를 채취합니다

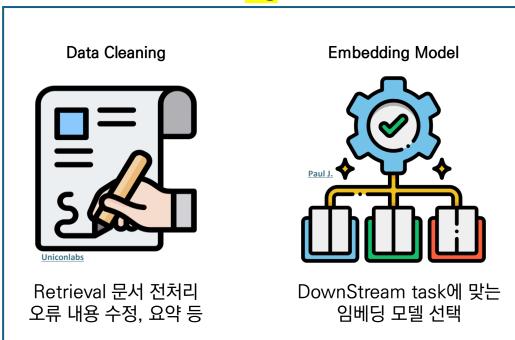
RAG를 통해 할루시네이션이 줄어들었으며, 리더보드 0.67에서 0.73으로 수직 상승 정성평가 결과, 다중질문에 대한 답변을 못하는 케이스를 고도화하기 위해 성능 높이는 RAG 기법 실험 진행

〈모범 답안〉

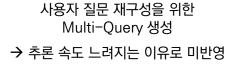
01. RAG - 고도화 실험

- 복잡한 질문이나 다중 질문에도 대답할 수 있게 모델이 참조하는 문서를 정제하고, 질문과 문서의 관련성을 높이는 것이 필요성이 도출됨
- 4가지 기법을 실험 후 Retrieval 문서 적합성, 추론 속도 등을 기준으로 Data Cleaning, Embedding Model 변경 방법을 최종 반영함

반영









후보 문서 관련성 기준 순위 재정렬 → 문서 적합성이 떨어져 미반영

01. RAG - Data Cleaning

확보한 데이터를 RAG 검색 문서로 사용하고 사용자의 질문과 관련성을 높이기 위해 전처리 작업을 수행해 <mark>리더보드 점수를 0.79로 향상시킴</mark>

(1) Train 데이터 정제

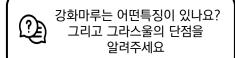
- 질문과 답변의 내용이 유사한 데이터가 존재하며, 서로 상충되는 내용을 가진 데이터도 있어 올바른 context를 찾는데 어려움이 있음
- Query에 대한 대답을 제대로 하기 위해, 유사한 내용의 질문-답변을 키워드 중심으로 그룹화 해 하나의 문서로 요약하는 작업을 수행함
- 사용자의 질문과 문서의 유사도를 높이기 위해 해당 문서와 관련된 질문을 문서에 추가했을 때 더 검색이 잘되는 효과가 있었음

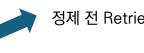
(2) Blog 데이터 정제

- Train 데이터에 없는 질문에도 잘 대답할 수 있도록 도메인 관련 지식을 보완하고자 함
- 이를 위해, 도배하자 및 인테리어 관련 블로그 글 스크랩한 후 Chat 형태의 답변으로 재구성한 상태에서, (1) 실험에서의 결과를 반영해 관련 질문을 문서에 포함시키는 작업 진행함

Train data

키워드	답변
강화마루	강화마루는 원목 무늬 ~
강화마루	강화마루의 단점은 열전도율이 ~
강화마루	강화마루의 기대수명은 ~
그라스울	그라스울의 단점은~
그라스울	그라수을의 장점은 ~





정제 전 Retrieval 문서:







강화마루 강화마루

강화마루

다양화







강화마루

그라스울

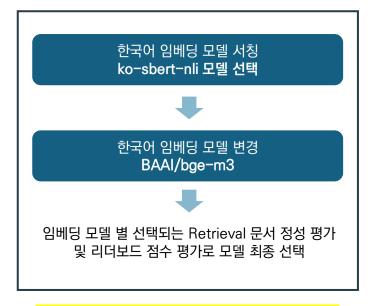
바닥재종류

정제 후 Retrieval 문서

01. RAG - 임베딩 모델 변경

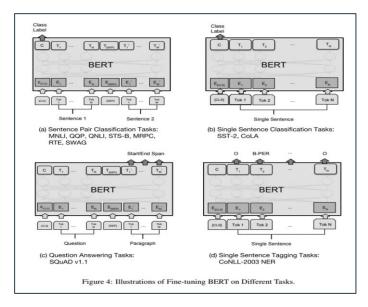
한국어 질문과의 유사도 파악을 위해 한국어 임베딩 모델을 변경하며 Retrieval 문서에 대한 적합성 평가와 리더보드 점수 기준으로 임베딩 모델 최종 선택함. Downstream task를 위한 추가 미세조정을 실행했으나 변화가 없어 오픈 소스로 공개된 BAAI/bge-m3 적용함

〈 한국어 임베딩 모델 변경 〉



리더보드 점수 0.73 → 0.76 (0.03 상승)

〈임베딩 모델 Fine-Tuning〉

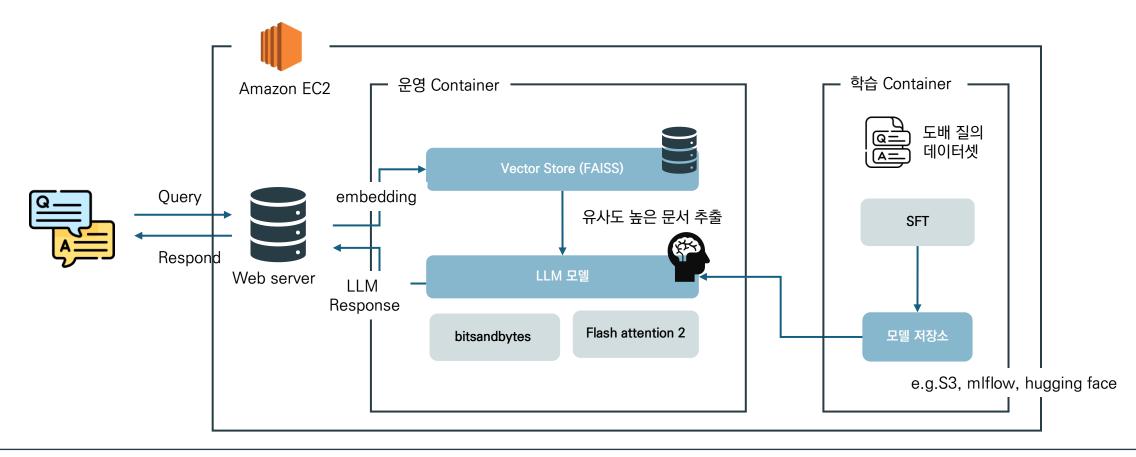


리더보드 점수 변동 없음 (적용 안함)

- 01. 대회 개요 및 대회 목적
- 02. 데이터 활용
- 03. 평가 전략
- 04. LLM 모델
- 05. RAG
- 06. 결론
- + Appendix

모델 서빙 프로세스 제안

- 사용량 변화에 따라 유연하게 Scale in/out 가능하도록 도커 컨테이너로 구성하고, 학습 시 GPU 자원 소모가 심한 학습 서버와 추론 서버 분리.
- 적은 비용으로 많은 사용자의 요청을 빠르게 처리할 수 있도록 추론 속도 빠른 모델(Orion) 선택.
- 메모리 효율성은 높여 빠른 속도를 제공하는 양자화 기법 라이브러리 bitsandbytes과, flash-attention2 적용



결론

DATA 질문-답변 맵핑 전수 조사 및 오탈자 검수를 통한 정확도 UP!



SFT

문제를 분석 및 정의하고 정확한 답변 도출을 위해 다양한 실험을 통한 정확도 & 속도 UP!



RAG

데이터를 키워드 중심으로 요약하고 외부데이터 수집을 통한 정확도 & 속도 UP! 리더보드 점수가 훨씬 좋은 모델도 존재하였으나,

"당장 사용 가능한 챗봇"에 부합한 LLM 모델 최종 선택

130개 추론 속도 : 약 8분 50초(GPU 사용 기준)

개당 추론 속도 : 약 4.08 초

대회 최종 Private 점수: 0.75842 / 상위 5%







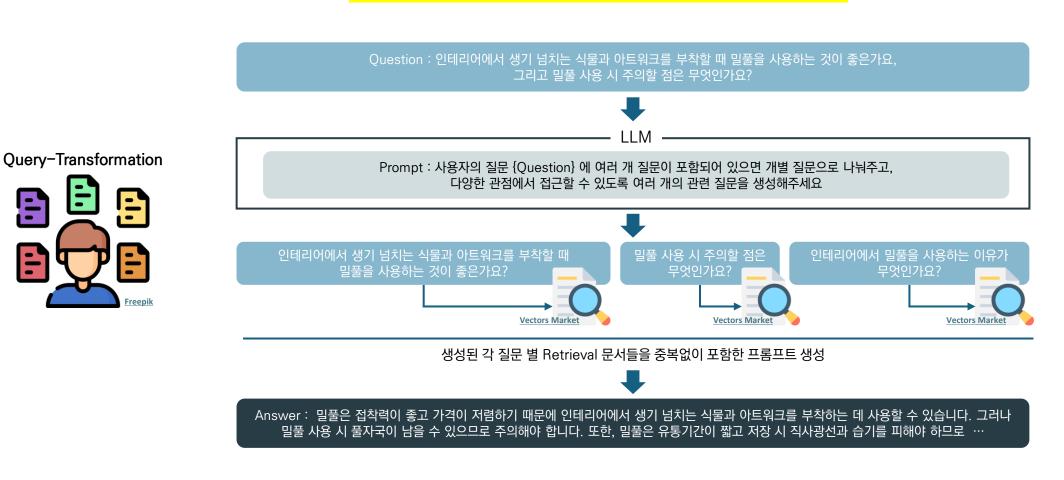


0.75842

- 01. 대회 개요 및 대회 목적
- 02. 데이터 활용
- 03. 평가 전략
- 04. LLM 모델
- 05. RAG
- 06. 결론
- + Appendix

RAG 문서 선택 정교화 기법 - Multi Query

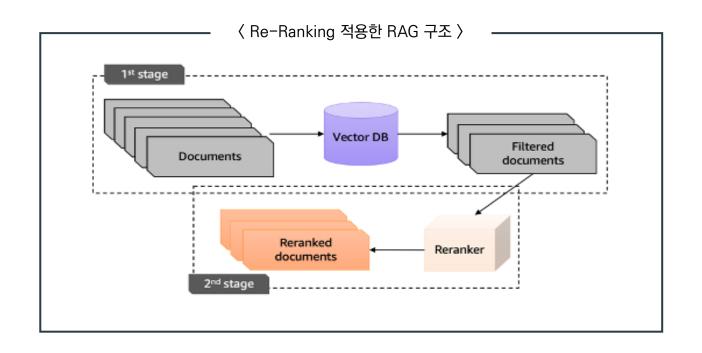
질문에 관련된 내용을 검색하기 어려운 다중 질문, 복잡한 질문 등을 해결하기 위해 LLM을 사용해 질문을 재구성하는 Multi Query 기법을 적용함. 실제로 다중 질문에 대한 답변 정확도는 향상되었으나 <mark>추론 속도가 130개 테스트데이터 기준 약 5분 늘어나 반영하진 않음</mark>



RAG 문서 선택 정교화 기법 - Re Ranking

Retrieval 문서는 질문과의 유사도를 기준으로 선택되기 때문에 관련성과 완전히 일치하지 않을 수 있는데, 이에 대해 후보 문서들에 대해 질문에 대한 관련성 및 일관성을 판단하여 문서의 우선 순위를 재정렬하는 기법인 Re-Ranking 적용했으나 답변 완성도가 떨어지는 문제가 있어 반영하지 않음.





Appendix

개발환경

GPU RTX3090* 2개

CUDA Version 11.8

Python 3.8.10

transformers==4.37.2

torch==2.1.0+cu118

accelerate==0.24.1

trl = 0.7.11

bitsandbytes==0.42.0

감사합니다