

The background of the slide is a grayscale photograph of a hand using a paint roller to apply dark paint to a light-colored wall. The roller is positioned on the right side, and the paint is being applied in a vertical stroke. The wall has some texture and shadows.

Dacon

도배하자 응답 Chat Bot

팀: XwenA

팀원: 최재빈, 서영우, 주용규, 윤성민

목차

0. 대회 소개

1. 데이터 수집 및 전처리

2. 모델 선정과 랭체인 구조 소개

3 . Fine tuning

4. 프로젝트 고찰



0. 대회 소개



0. 대회 소개 - 내용

도배 하자 질의 응답 처리 : 한솔데코 시즌2 AI 경진대회

알고리즘 | 언어 | LLM | MLOps | QA | Cosine Similarity

₩ 상금 : 1000만 원

🕒 2024.01.29 ~ 2024.03.11 09:59

[+ Google Calendar](#)

👤 1,310명 🗓 마감

[주제]

도배 하자 질의 응답 AI 모델 개발

[설명]

도배 하자 도메인에 대한 질의를 바탕으로 지능적인 응답을 생성하는 AI 모델 개발해야 합니다.

[총 상금 1,000만 원]

🥇 1위 - 600만 원

🥈 2위 - 300만 원

🥉 3위 - 100만 원



0. 대회 소개 - 데이터 처리 방법

train.csv test.csv sample_submission.csv ▾						
Views Grid view Hide fields Filter Group Sort						
<input type="checkbox"/>	id	질문_1	질문_2	category	답변_1	답변_2
1	TRAIN_000	면진장치가 뭐야?	면진장치에 사용되는 주요 기	건축구조	면진장치란 지반에서 오는 진	면진장치란 건물의 지반에서
2	TRAIN_001	내진설계의 종류 좀 알려줘	내진설계에는 어떤 종류가 있	건축구조	내진 설계의 종류로 내진구조	내진설계에는 내진구조, 제진
3	TRAIN_002	철골구조의 장점이 뭐야?	철골구조의 장점을 알려줘?	건축구조	철골구조는 건물의 외벽에	철골구조의 장점은 건물의 외
4	TRAIN_003	철골철근 콘크리트 구조가 뭐	철골철근 콘크리트 구조의 장	건축구조	철근철골콘크리트는 철골과	철골철근콘크리트 구조는 건
5	TRAIN_004	철골구조는 어떤 방식이 있어	철골구조의 다양한 방식이 무	건축구조	철골구조는 일반철골구조와	철골구조는 일반철골구조와

train.csv test.csv sample_submission.csv ▾		
Views Grid view Hide fields Filter Group Sort		
<input type="checkbox"/>	id	질문
1	TEST_000	방정 폐인트의 종류에는 어떤 것들이 있는지 알고 계신가요? 또한, 원목사이딩을 사용하는 것에 어떤 단점이 있을까요?
2	TEST_001	도배지에 녹은 자국이 발생하는 주된 원인과 그 해결 방법은 무엇인가요?
3	TEST_002	큐블럭의 단점을 알려주세요. 또한, 압출법 단열판을 사용하는 것의 장점은 무엇인가요?
4	TEST_003	철골구조를 사용하는 고층 건물에서, 단열 효과를 높이기 위한 시공 방법은 무엇이 있을까요?
5	TEST_004	도배지의 완전한 건조를 위해 몇 주 동안 기다려야 하나요?

모델 학습과 추론에서 평가 데이터셋 정보 활용(Data Leakage)시 수상 제외 (평가 데이터셋의 Pseudo Labeling 포함)

모든 학습, 추론의 과정 그리고 추론의 결과물들은 정상적인 코드를 바탕으로 이루어져야하며, 비정상적인 방법으로 얻은 제출물들은 적발 시 규칙 위반에 해당됩니다.

- 정상적인 코드의 범주 : 반드시 제공 데이터로부터 언어 모델 학습의 과정이 존재해야하며, 학습된 언어 모델을 바탕으로 추론이 이루어져야합니다.



0. 대회 소개 - 답변 평가 방법

평가 산식 : Cosine Similarity

Encode Model : distiluse-base-multilingual-cased-v1

id	vec_0	vec_1	vec_2	vec_3	vec_4	vec_5	vec_6	vec_7	vec_8	vec_9	vec_10
TEST_000	-0.01596	0.067004	0.032767	-0.00498	0.066003	-0.01307	0.026254	0.060146	0.000336	0.013178	-0.09845
TEST_001	-0.01945	0.018962	-0.01039	0.024137	0.082303	-0.02705	-0.01499	-0.03579	0.000561	0.010709	-0.11011
TEST_002	-0.01319	-0.04331	-0.04563	-0.0104	0.121919	-0.0381	0.006298	-0.01464	0.069159	0.046603	-0.00709
TEST_003	0.000612	-0.00928	0.019282	0.005767	0.012939	-0.07395	-0.0656	-0.01138	0.007521	0.000121	0.068324
TEST_004	0.013479	0.001879	-0.01375	-0.0276	0.114445	-0.02803	0.064021	0.054996	-0.01243	-0.00192	-0.01329
TEST_005	0.014341	0.005115	0.007307	-0.00123	0.067792	-0.03008	-0.07317	-0.01581	0.023657	0.01504	0.023201
TEST_006	-0.03517	0.019673	-0.01077	-0.00012	0.114836	-0.02389	-0.02968	0.027813	-0.0062	0.026484	0.012949
TEST_007	-0.0136	-0.00473	-0.03438	0.009285	0.104334	0.031743	0.071624	0.028873	0.008315	-0.00493	-0.03831
TEST_008	-0.01672	1.47E-05	-0.03537	-7.75E-05	0.035636	-0.00275	0.06934	-0.0224	0.004346	0.020348	-0.04967
TEST_009	0.009143	-0.01239	0.003188	0.031045	0.040253	-0.02606	-0.07162	0.004672	-0.02259	0.01985	-0.011
TEST_010	-0.04176	-0.00228	-0.05743	0.02046	0.081974	0.004487	0.052887	-0.01855	-0.005	0.078518	-0.10339



A person with long brown hair in a ponytail, wearing a dark jacket, is seen from the back, painting a large, irregular black shape on a light-colored, textured wall. The person is holding a paint roller. The background wall has a rough, stone-like texture with some circular patterns. The title text is centered within the black painted area.

1. 데이터 수집 및 전처리

1. 데이터 수집 및 전처리 - 데이터 수집

◎ 데이터 질문 범위 확대

- train.csv 질문 증강 (구글 시트 GPT 확장기능 사용)



질문1 생성 프롬프트	질문1 생성	질문1 좋은것만 뽑은거
아래의 질문과 카테고리를 참고하여 해당 카테고리에 맞는 완전히 새로운 질문을 창작 해줘 질문: 콘크리트 구조는 어떤 방식이 있어? 카테고리: 건축구조 새로운 질문:	건축구조에서 사용되는 콘크리트의 다양한 형태는 무엇인가요?	철골철근 콘크리트 구조의 장점과 단점은 무엇인가요?

1. 데이터 수집 및 전처리 - 데이터 수집

◎ 구글 제미나이 api 를 활용하여 데이터 증강

```
import os

os.environ['GOOGLE_API_KEY'] = 'AIzaSyCEwhUeEALqnHfzMC9Xlvm53kqErg1hGc4'
GOOGLE_API_KEY = os.getenv('GOOGLE_API_KEY')
genai.configure(api_key=GOOGLE_API_KEY)
```

Processing Questions: 1% | 1/100 [00:43<1:12:17, 43.82s/it] 질문: 거실 분위기를 한층 더 돋보이게 하는 커튼 소재는 무엇인가요?

답변: 실크, 벨벳, 브로케이드와 같은 고급스러운 소재를 고려해 보세요. 이러한 소재는 빛을 반사하여 공간에 따뜻함과 우아함을 더합니다.

Processing Questions: 2% | 2/100 [01:26<1:10:37, 43.24s/it] 가장 중요한 점은 효율성입니다.

Processing Questions: 3% | 3/100 [02:10<1:10:24, 43.55s/it]

아래카 야자의 잎은 실내 공간에 열대적 분위기를 더해주는 장식 용품으로 활용할 수 있습니다. 잎을 화병에 꽂아 싱싱한 분위기를 연출하거나, 건조시킨 잎을 프레임

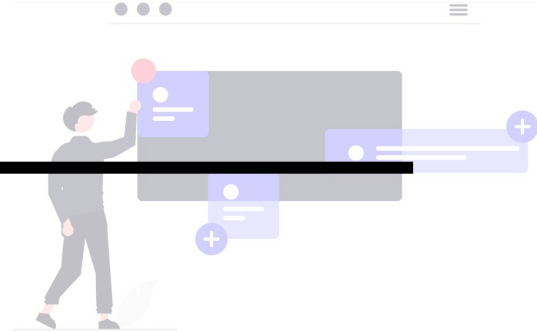
Processing Questions: 4% | 4/100 [02:54<1:10:11, 43.87s/it] 인테리어 디자인에서 물딩을 천장과 벽의 결합부에 설치하는 것은 다음과 같은 이유로 중

- 실내에 우아하고 정교한 분위기를 더해줍니다.
- 천장과 벽 사이의 결합이나 불규칙성을 숨겨줍니다.
- 공간에 구조적 깊이와 차원감을 더해줍니다.
- 그림 프레임이나 벽걸이와 같은 다른 장식 요소를 위한 배경을 제공합니다.

Processing Questions: 5% | 5/100 [03:39<1:10:06, 44.27s/it] 작은 베란다는 야외 라운지로 변신시키는 데 필요한 주요 디자인 요소는 다음과 같습니다.

- * 가구: 편안하고 날씨에 견디는 가구를 선택하세요.
- * 텍스타일: 방수 소재의 쿠션과 담요를 사용하여 편안함과 스타일을 더하세요.
- * 조명: 분위기를 조성하고 안전성을 높이려면 조명을 추가하세요.
- * 식물: 베란다는 생명력을 더하고 프라이버시를 제공하는 식물을 배치하세요.
- * 액세서리: 베란다는 개성을 더하는 액세서리, 예를 들어 화분, 양초, 투척 베개를 추가하세요.

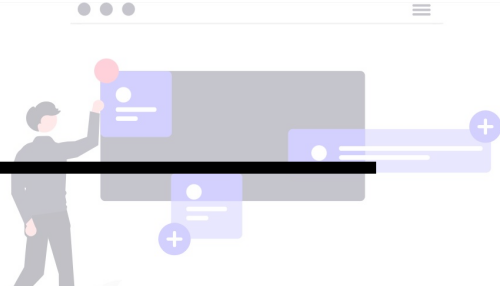
1. 데이터 수집 및 전처리 - 데이터 수집



✓ 도배 하자 레퍼런스_데이콘제공

- 1. 공동주택에서 하자담보책임과 하자보수보증책임에 관한 연구.pdf
 - 2. 신축공동주택 실내마감공사 하자 원인분석 및 예방방안.pdf
 - 3. 시대별 공동주택 설계변천에 따른 하자유형 변화 및 특징에 관한 연구.pdf
 - 4. 벽지반점_조사연구_내화건축자재협회.pdf
 - 5. 공동주택타일공사의하자유형분석및저감대책에관한연구.pdf
 - 6. 공동주택의 목질계 바닥 마감계의 하자 유형 분석 및 대책.pdf
 - 7. 공동주택의 공정별 하자발생 및 예방 대책에 관한 연구.pdf
 - 8. 공동주택 하자발생 원인과 방지방안에 관한 연구.pdf
 - 9. 공동주택 세대내부 결로 발생 하자 방지를 위한 단열재 시공 개선방안에 관한 연구.pdf
 - 10. 공동주택 건축공사 하자유형별 중요도 분석.pdf
 - 11. 공동주택 하자판정기준 주요 신설사항 소개.pdf
 - 12. CNN기반 공동주택 도배하자 검출 딥러닝 이미지 분류모델 구축에 관한 연구.pdf
 - 13. 공동주택 건축공사의 하자유형과 대책에 관한 연구.pdf
- ≡ 논문순위.txt
- ≡ 논문요약.txt

1. 데이터 수집 및 전처리 - 데이터 수집



```
1 train_add = pd.read_csv("/content/gdrive/MyDrive/kdt_jyg/workspace/Dacon_hansol/retriever_test/Q_A_240305.csv", encoding='utf-8')
2 for index, row in train_add.iterrows():
3     question = row['Q']
4     answer = row['A']
5     train_data.append("질문: " + question + " 답변: " + answer)
```

```
1 print(len(train_data))
```

8136

```
1 print(train_data[7001])
```

질문: 건물 구조물에서 샤프가 가장 중요한 기능은 무엇인가요? 답변: 건물 구조물에서 샤프의 가장 중요한 기능은 외부 요소로부터 실내를 보호

“질문: ... 무엇인가요? 답변: ... 입니다.” 형식의 데이터로 가공하여 사용.

2. 모델 선정과 랭체인 구조 소개



2. 모델 선정과 랭체인 구조 소개 – 42dot 모델 선택

◎ 42dot 모델 선택 이유

- 학습 데이터셋이 유사한 구조 (질문/요청 및 이에 대한 응답으로 이루어진 Single/Multi-turn 형태)
- 모델의 크기가 다른 Llama모델에 비해 가벼움 (하단 자료 참고)

42dot LLM-SFT (생성형 언어 모델)

42dot LLM-SFT는 42dot LLM-PLM에 SFT (Supervised Fine-Tuning)를 수행한 모델로, 학습을 위한 파라미터는 아래와 같습니다.

Model	Global Batch Size	Learning rate	Epochs	Max length	Weight decay	Warmup ratio
1.3B	16	2e-5	3	4,096	0	0.03

학습은 A100 80G GPU 16장을 사용했고 약 7시간 (112 GPU hours) 소요되었습니다.

학습 데이터셋

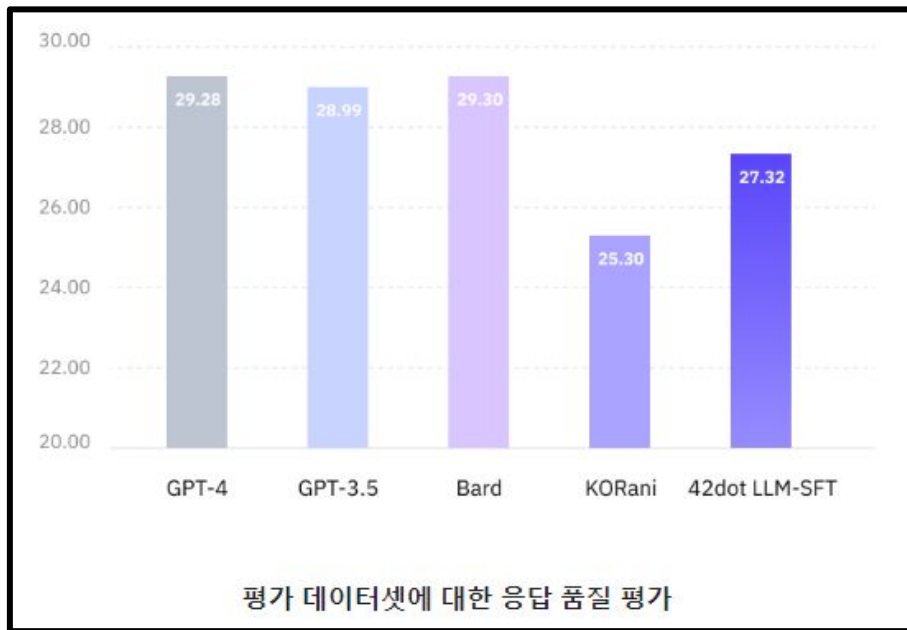
질문/요청 및 이에 대한 응답으로 이루어진 Single/Multi-turn 형태의 대화 데이터를 학습에 사용했습니다.



2. 모델 선정과 랭체인 구조 소개 - 모델 평가 내역

◎ 42dot 모델 평가내역

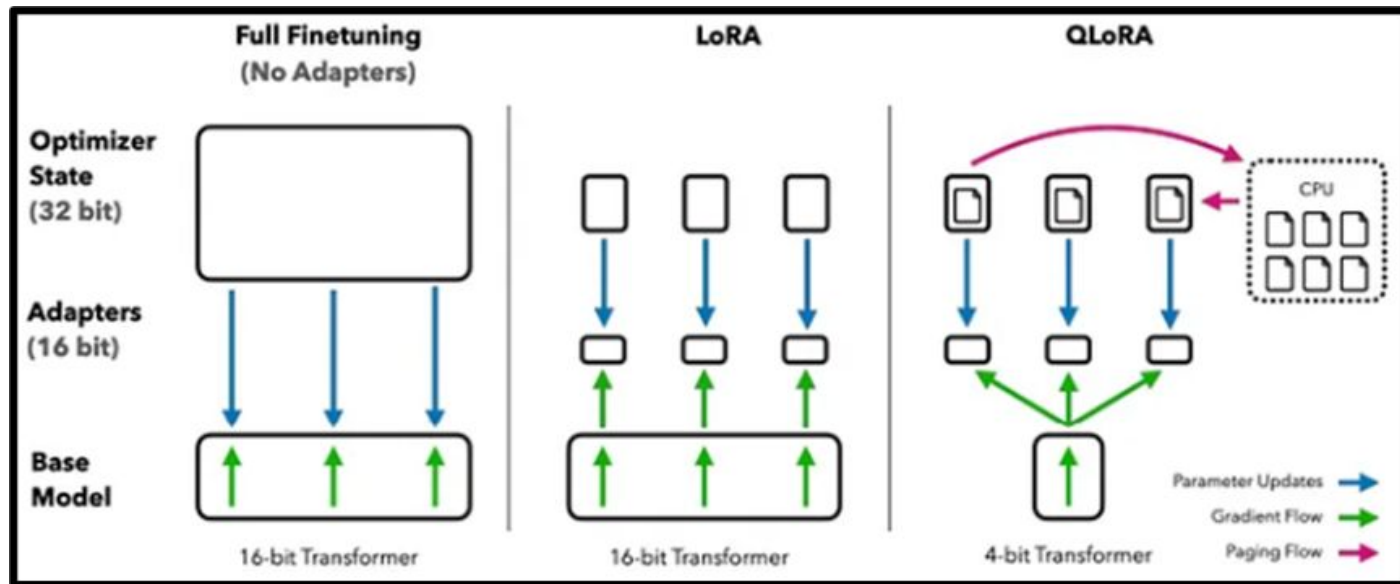
- Accuracy, Robustness, Fairness, Bias, Toxicity, Efficiency 각 5점으로 평가한 점수



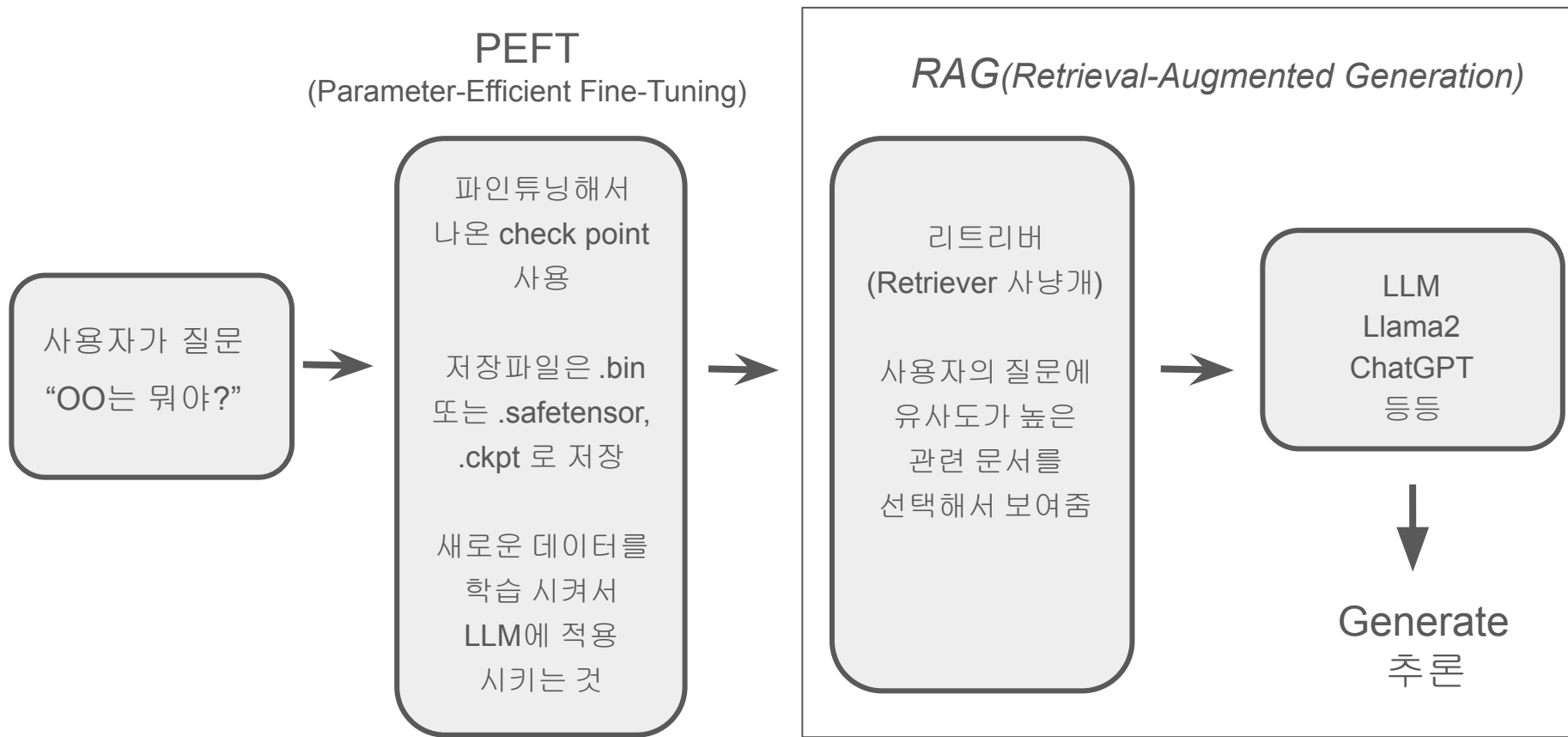
2. 모델 선정과 랭체인 구조 소개 – QLoRA

© QLoRA(Quantized Low Rank Adapters)

- 이 방법은 4비트로 양자화된 베이스 모델에 LoRA 어댑터를 추가하여 메모리 사용량을 크게 줄이고, 16비트 풀 파인튜닝 성능을 유지합니다.



2. 모델 선정과 랭체인 구조 소개 - 랭체인 구조



3. Fine tuning



3. Fine tuning – 문제 발생 분석

- 같은 단어를 반복 사용 하여 토큰을 낭비하는 경우
- 끝맺음이 원활하지 않은 경우
- 답변을 제대로 하지 못하는 문제

2%|| | 2/130 [00:16<17:54, 8.39s/it]
질문: 도배지에 녹은 자국이 발생하는 주된 원인과 그 해
답변: 도배지에 녹은 자국이 발생하는 주요 원인은 다음과
1. 도배지의 접착제 층이 너무 얇거나 균일하지 않을 때
2. 도배지의 접착제 층에 균열이 있을 때
3. 도배지의 접착제 층에 결로가 있을 때
4. 도배지의 접착제 층에 습기가 있을 때
5. 도배지의 접착제 층에 곰팡이가 있을 때
6. 도배지의 접착제 층에 먼지나 이물질이 있을 때
7. 도배지의 접착제 층에 수분이 너무 많을 때
8. 도배지의 접착제 층에 압력이 너무 높을 때
9. 도배지의 접착제 층에 열이 너무 높을 때
10. 도배지의 접착제 층에 습기가 너무 많을 때

30%|| | 39/130 [03:17<09:22, 6.18s/it]
질문: 방청도료 도장 작업을 위해 필요한 단계는 무엇인가? 또한, 콘크리트
답변: 방청도료를 도장하기 위해 필요한 단계를 알려드리겠습니다.
1. 방청 도료를 선택합니다. 방수 도료와 방청 페인트를 혼합하여 사용하는 것
2. 도장할 벽의 표면을 깨끗하게 청소합니다. 먼지와 오염물을 제거하여 도장
3. 도료의 두께를 조절하기 위해 도장 도구를 선택합니다 (예: 브러시, 롤러,
4. 도료를 도포하기 전에 벽의 결함을 확인합니다. 결함이 있는 부분은 도료를
5. 도료가 건조될 때까지 기다린 후, 벽에 도료를 바릅니다. 도막이 균일하고
6. 도장이 마른 후, 도막을 보호하기 위해 보호 필름을 사용합니다. 보호 필름
콘크리트를 뚫는 데 사용되는 도구와 기술은 다양합니다. 일반적으로 사용되는
1. 드릴: 콘크리트를 뚫을 때 사용되는 도구로, 구멍을 천공하는 데 사용됨
2. 드릴 비트: 드릴의 구멍을 더 깊게 뚫을 수 있도록 도와주는 도구로 사용됨
3. 드릴용 액세서리: 드릴을 보호하고 구멍을 균일한 크기로 뚫을 수 있는 도
4. 드릴 보호제: 구멍을 보호하는

45%|| | 58/130 [06:14<06:45, 5.64s/
질문: 강마루 바닥재의 장점은 무엇인가?
답변: 강마루는 어떤 종류의 바닥재인가?
강마루는 어떤 특징을 가지고 있나요?
어떤 상황에서 강마루를 사용하는 것이 좋을까요?
어떻게 강마루의 단점을 보완할 수 있을까요?
답변을 부탁드립니다.

3. Fine tuning – 적용 방법 소개

0. 데이터 분류 및 적용 방법

- 기본 train.csv 데이터 (6440개의 데이터 산출)
- 증강 train.csv 데이터 (증강질문 + GPT 답변. 1696개의 데이터 산출)
- 다중질문 train.csv 데이터 (다중 질문 형식의 데이터 1288개 산출)
- ※ 기타방법: 증강방식2만 적용. 답변형태의 데이터만 적용.

1. Langchain 모델 훈련 방법

- 상단의 데이터를 QLora 방식으로 훈련
- Parameter 조정, Epoch 횟수 조정, Merge 방식 조정 등등

2. RAG Chain 생성 방법

- 기본적으로 FAISS방식의 Retriever 사용.
- EnsembleRetriever 방식의 Retriever 사용. (FAISS 방식과 혼합하여 사용.)
- bm25, TFID, Ensemble 방식, Prompt 변경 등등 다양한 방식을 시도.

자세한 내용은
데모 시연 및
코드리뷰에서 확인

3. Fine tuning – 최종 적용 방법

Data Set

◇ Lang Chain 모델 훈련

- 기본 train.csv 파일

◇ Rag Chain 적용 Data

- 기본 train.csv 파일
- 증강 Data
 1. 기존 train 파일의 질문 증강
 2. 신규 질문 및 답변 추가

Lang Chain

◇ 사용 모델

- 42dot_LLM-SFT-1.3B

◇ 훈련 방식

- Qlora 방식 적용

◇ Epoch 방식

- 5회 훈련 후 merge 모델에
추가로 5회 Epoch 적용

Rag Chain

◇ Pipeline

- Langchain 훈련모델 적용

◇ Retriever 방식

- Ensemble 방식의 혼합방식

◇ Retriever 적용내역

- FAISS Retriever 가중치 0.5
- BM25 Retriever 가중치 0.5

4. 프로젝트 결과






















4. 프로젝트 고찰 - 제출로그

날짜	점수	적용 모델	모델 데이터	적용 리트리버	ret 데이터	epoch	적용 내역	특이사항
2024-02-13	0.6674	beam/llama-2-ko-7b	train.csv	FAISS	train.csv	0	일명 천사코드. (데이콘 공유 자료)	langchain 및 rag_chain 코드값 기본 베이스
2024-02-13	0.5653	skt/kogpt2-base-v2	train.csv	없음	train.csv	0	데이콘 제공 베이스라인 모델	
2024-02-17	0.4704	skt/kogpt2-base-v2	train.csv	없음	train.csv	0	max token: 512 / temperature: 0.3	훈련된 모델을 직접 파인튜닝할 경우 점수 하락
2024-02-23	변경 내역: 42dot 모델 적용 / QLora 방식의 파인튜닝 적용							
2024-02-23	0.6121	42dot_LLM-SFT-1.3B	train.csv	없음	train.csv	1		
2024-02-23	0.5967	42dot_LLM-SFT-1.3B	train.csv	없음	train.csv	1	temperature: 0.3 / prompt 300자	
2024-02-24	0.6460	42dot_LLM-SFT-1.3B	train.csv	없음	train.csv	5		
2024-02-24	0.6417	42dot_LLM-SFT-1.3B	train.csv	없음	train.csv	5	prompt 300자	
2024-02-24	0.6344	42dot_LLM-SFT-1.3B	train.csv	없음	train.csv	5	prompt 300자 / temperature: 0.2 / no_repeat: 3	
2024-02-25	0.6492	42dot_LLM-SFT-1.3B	train.csv	없음	train.csv	10	max_length: 512 / temperature 0.3 / no_repeat: 0	
2024-02-25	0.6243	42dot_LLM-SFT-1.3B	train.csv	없음	train.csv	10	temperature 0 / no_repeat: 3	
2024-02-25	0.6209	42dot_LLM-SFT-1.3B	train.csv	없음	train.csv	10	max_length: 600 / temperature 0.3 / no_repeat: 3	
2024-02-26	변경 내역: RAG_CHAIN 적용							
2024-02-26	0.7130	42dot_LLM-SFT-1.3B	train.csv	FAISS	train.csv	0	RAG_CHAIN 첫 적용	기본모델로 테스트
2024-02-27	0.7158	42dot_LLM-SFT-1.3B	train.csv	FAISS				
2024-02-27	0.7143	42dot_LLM-SFT-1.3B	train.csv	FAISS				
2024-02-28	0.7212	42dot_LLM-SFT-1.3B	train.csv	FAISS				
2024-02-28	0.7088	42dot_LLM-SFT-1.3B	train.csv	FAISS				
2024-02-29	0.7001	42dot_LLM-SFT-1.3B	train.csv	FAISS				
2024-02-29	0.7056	42dot_LLM-SFT-1.3B	train.csv	FAISS				
2024-03-04	적용 내역: Best 모델 Prompt 적용 (점수 비교 및 검증)							
2024-03-04	0.7020	42dot_LLM-SFT-1.3B	train.csv	bm25	train.csv	5+5	기존 Best 모델, bm25 리트리버 적용	
2024-03-04	0.7271	42dot_LLM-SFT-1.3B	train.csv	Ensemble	train.csv	5+5	기존 Best 모델, 앙상블 리트리버 적용	
2024-03-05	0.6796	42dot_LLM-SFT-1.3B	QA.csv	Ensemble	train.csv	10	증강질문 훈련모델, 앙상블 리트리버	gpt, gemini 증강질문 1600개
2024-03-05	0.6394	42dot_LLM-SFT-1.3B	QA.csv	FAISS	train.csv	10	증강질문 훈련모델, 페이스 리트리버	gpt, gemini 증강질문 1600개
2024-03-05	0.7380	42dot_LLM-SFT-1.3B	train.csv	Ensemble	train + QA	5+5	기존 Best 모델, 앙상블 리트리버, 리트리버 데이터 증강	리트리버 데이터 증강효과 확인
2024-03-06	적용 내역: Best 모델 Prompt 적용, Ensemble 리트리버 적용, 리트리버 데이터에 증강 적용 (train + QA = 8136개)							
2024-03-06	0.7162	42dot_LLM-SFT-1.3B	Q2.csv	Ensemble	train + QA	10	증강 질문2만 사용하여 훈련 (GPT 답변 비중 높음)	
2024-03-06	0.7140	42dot_LLM-SFT-1.3B	train + QA	Ensemble	train + QA	6	기존+증강 데이터를 사용하여 훈련	
2024-03-06	0.7143	42dot_LLM-SFT-1.3B	train.csv	Ensemble	train + QA	5+5	리트리버 데이터의 답변만을 사용	
2024-03-07	0.7123	42dot_LLM-SFT-1.3B	Q2.csv	Ensemble	train + QA	5+5	증강 질문2만 사용 및 5+5방식의 merge 훈련	
2024-03-07	0.7142	42dot_LLM-SFT-1.3B	train + QA	Ensemble	train + QA	10	기존+증강 데이터를 사용하여 훈련	
2024-03-07	0.7271	42dot_LLM-SFT-1.3B	train.csv	Ensemble	train + QA	5+5	0.738 테스트 조건에서 prompt를 영어로 변경 및 500자 제한	
2024-03-08	0.7184	42dot_LLM-SFT-1.3B	train.csv	Ensemble	train + QA	5+5	Best 5+5 훈련조건과 동일 훈련모델의 점수를 확인	738 모델은 코업으로 훈련, loss값이 다르게 나옴
2024-03-08	0.7268	42dot_LLM-SFT-1.3B	train.csv	Ensemble	train + QA	5+5	리트리버 1+1 앙상블 페이스 가중치 0.7	페이스 리트리버 보다 점수 낮음
2024-03-08	0.7310	42dot_LLM-SFT-1.3B	train.csv	Ensemble	train + QA	5+5	유저 프롬프트 변경 (rag.stream 부분)	내용: 건축 전문가적 입장에서 300자 이내로 요약해서 답변해주세요.
2024-03-09	0.7200	42dot_LLM-SFT-1.3B	train.csv	Ensemble	train + QA	5+5	유저 프롬프트 변경 (rag.stream 부분)	내용: 요약 단답형 어떻게 방법
2024-03-09	0.7186	42dot_LLM-SFT-1.3B	train.csv	Ensemble	train + QA	5+5	리트리버 변경 (TFIDFRetriever)	
2024-03-09	0.7284	42dot_LLM-SFT-1.3B	1288 가공	Ensemble	train + QA	100	모델 훈련 케이스 변경_1288개 100epoch	
2024-03-10	0.6947	42dot_LLM-SFT-1.3B	train.csv	Ensemble	train + QA	5+5	유저 프롬프트 변경 (rag.stream 부분)	내용: 요약 단답형 정확히
2024-03-10	0.5895	42dot_LLM-SFT-1.3B	1288 다중질문	Ensemble	train + QA	50	50 다중질문 에포크	
2024-03-10	0.7297	42dot_LLM-SFT-1.3B	train.csv	Ensemble	train + QA + 다중	5+5	738에 다중질문 데이터 추가	
2024-03-11	0.7256	42dot_LLM-SFT-1.3B	train.csv	Ensemble	train + 다중	5+5	738에 다중질문 데이터 추가 (이전 증강데이터는 삭제)	

최저점수: 0.4704 → 최고점수 : 0.7380

4. 프로젝트 고찰 - 리더보드 점수 (최종 점수 성적)

67	kaffeine		0.73385	15	21일 전
68	호호호호호		0.73349	5	3일 전
69	jln21		0.73293	4	12일 전
70	steelbear	 	0.73247	22	25일 전
71	chapro		0.73226	3	25일 전
72	하이기니		0.73131	41	3일 전
73	진지잡썸		0.73119	8	9일 전
74	XwenA	   	0.73055	44	7일 전
75	knlpscience	  	0.72858	18	하루 전
76	덩먹	 	0.72836	7	한 달 전
77	GomTaengYi	 	0.72778	21	8일 전

SS
CUSTOMER BUSINESS
WORK
PERFORMANCE
TEAM
SUPPORT
STRATEGY
GOALS

PPT 발표를 끝났습니다.
프로그램 시연을 진행하겠습니다.
감사합니다.

피드백 - RAG 사용 전 Model 답변 생성

3. Fine tuning – 모델 파라미터 수정

model.generate 의 파라미터를 수정 (RAG 체인 적용 전)

```
generated_ids = model.generate(  
    tokens["input_ids"],          # 모델에 입력할 토큰 ID들입니다. 이전에 토큰화한  
    max_length=400,              # 생성할 텍스트의 최대 길이를 지정합니다. 이 경우 최대 3  
    min_length=50,               # 생성할 텍스트의 최소 길이를 지정합니다. 적어도  
    num_return_sequences=1,      # 생성할 텍스트 시퀀스의 수를 지정합니다. 여기서는 1개의  
    pad_token_id=tokenizer.eos_token_id, # 패딩에 사용할 토큰 ID를 지정합니다. 여기  
    eos_token_id=tokenizer.eos_token_id, # 문장 종료 토큰의 ID를 지정합니다. 생성된  
    no_repeat_ngram_size=3,      # 반복되는 n-gram의 크기를 지정합니다. 여기서는 3.  
    temperature=0.2,             # 생성 과정에서의 무작위성 정도를 지정합니다.  
    # top_p=0.6,                 # 0과 1 사이의 값을 설정합니다. 0에 가까울수록 가장 높은 확률을 가  
    # repetition_penalty=4.8     # 0 이상의 값을 설정합니다. 값이 높을수록 동일한 단어 또  
)
```

※ 적용 결과

→ 0.59 ~ 0.64 정도의 점수가 출력됨.

(max_length: 512 / temperature 0.3 / no_repeat: 0 적용시 최고점. 자세한 내용은 제출로그 Page 참고)

프롬프트 적용 방식

3. Fine tuning – Prompt 엔지니어링 (SYSTEM)

– RAG체인에서 데이터 추출에 사용되는 Custom Prompt를 수정

```
from langchain_core.prompts import PromptTemplate
from langchain_core.runnables import RunnablePassthrough
from langchain_core.output_parsers import StrOutputParser

def format_docs(docs):
    return "\n\n".join(doc.page_content for doc in docs)

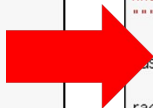
template = """건축 전문가적 입장에서 아래 맥락을 확인하고 질문에 반드시 300자 이내의 답변을 해줘 답변은 반드시 한국어로 해줘
맥락: {context}

질문: {question}

답변:"""

custom_rag_prompt = PromptTemplate.from_template(template)

rag_chain = (
    {"context": ensemble_retriever | format_docs, "question": RunnablePassthrough()}
    | custom_rag_prompt
    | hf
    | StrOutputParser()
)
```



```
from langchain_core.prompts import PromptTemplate
from langchain_core.runnables import RunnablePassthrough
from langchain_core.output_parsers import StrOutputParser

def format_docs(docs):
    return "\n\n".join(doc.page_content for doc in docs)

template = """From an architectural expert's perspective, please review the context below and
Context: {context}

Question: {question}

Answer:
"""

custom_rag_prompt = PromptTemplate.from_template(template)

rag_chain = (
    {"context": ensemble_retriever | format_docs, "question": RunnablePassthrough()}
    | custom_rag_prompt
    | hf
    | StrOutputParser()
)
```

※ Custom Prompt 영어 변경 결과

→ 점수 하락. 시스템 프롬프트는 기존의 프롬프트를 유지. (제출로그 Page 참고)

3. Fine-tuning 과 RAG – 프롬프트 엔지니어링 (USER)

- RAG체인에서 사용자의 질문이 입력되는 부분에 추가 명령어를 입력하여 결과값을 조정.

포세린 타일의 장점과 단점은 무엇인가요?

포세린 타일은 밀도와 강도가 뛰어나고 변형이 적으며 반영구적이라는 장점이 있습니다. 또한 내구성이 뛰어나고 흡수율이 낮아 물에 대해 저항력을 가지며 각종 화학물질 및 열에 대한 저항성이 뛰어나다는 장점을 가지고 있습니다. 그러나 시공 방법에 따라 하자 발생률이 높을 수 있으며 오염에 취약하다는 단점도 있습니다. 도배시에는 시공사의 전문적인 기술과 꼼꼼한 관리가 필요하며, 오랜 시간 사용할 경우에도 적절한 관리가 필요합니다.

질문: 포세린 타일의 장점과 단점은 무엇인가요?

답변: 포세린 타일은 밀도와 강도가 뛰어나고 변형이 적으며 반영구적이라는 장점이 있습니다. 또한 내구성이 뛰어나고 흡수율이 낮아 물에 대해 저항력을 가지며 각종 화학물질 및 열에 대한 저항성이 뛰어나다는 장점을 가지고 있습니다. 그러나 시공 방법에 따라 하자 발생률이 높을 수 있으며 오염에 취약하다는 단점도 있습니다. 외부 충격에도 강하고 세척 및 청소가 용이하며 다양한 디자인의 선택이 가능하여 실내, 외부 및 습기가 많은 환경에 모두 적합하다는 점도 장점으로 꼽힙니다.

3. Fine-tuning 과 RAG – 프롬프트 엔지니어링 (USER)

– 하단과 같이 기본질문 + 명령어 입력경우 답변이 변경됨을 확인.

포세린 타일의 장점과 단점은 무엇인가요? 요약 단답형 어떻게 방법

포세린 타일의 장점은 내구성이 뛰어나고 바닥 울림이 적다는 것입니다.

또한, 가격이 상대적으로 저렴하고 충격 흡수율이 높아 노약자가 있는 집안에도 적합합니다.

그러나 시공 방법에 따라 하자 발생률이 높을 수 있고 오염에 취약하다는 단점이 있습니다.

◇ 적용 결과

1. 건축 전문가적 입장에서 300자 이내로 요약해서 답변해주세요. → 점수 하락
2. 요약 단답형 어떻게 방법 → 점수 하락
3. 요약 단답형 정확히 → 점수 하락

※ 하락원인 분석: 답변이 길어야 될 부분도 줄여서 출력하여 Cosine Similarity가 하락한 것으로 예상.

모델 훈련 Case 별 점수

모델 훈련 횟수 별 점수

날짜	점수	적용 모델	모델 데이터	적용 리트리버	ret 데이터	epoch	적용 내역
2024-02-24	0.6460	42dot_LLM-SFT-1.3B	train.csv	없음	train.csv	5	
2024-02-24	0.6417	42dot_LLM-SFT-1.3B	train.csv	없음	train.csv	5	prompt 300자
2024-02-24	0.6344	42dot_LLM-SFT-1.3B	train.csv	없음	train.csv	5	prompt 300자 / temperature: 0.2 / no_repeat: 3
2024-02-25	0.6492	42dot_LLM-SFT-1.3B	train.csv	없음	train.csv	10	max_length: 512 / temperature 0.3 / no_repeat: 0
2024-02-25	0.6243	42dot_LLM-SFT-1.3B	train.csv	없음	train.csv	10	temperature 0 / no_repeat: 3
2024-02-25	0.6209	42dot_LLM-SFT-1.3B	train.csv	없음	train.csv	10	max_length: 600 / temperature 0.3 / no_repeat: 3
2024-02-28	0.7088	42dot_LLM-SFT-1.3B	train.csv	FAISS	train.csv	5+5+5	프롬프트 1000자 제한, epoch 분산
2024-02-29	0.7001	42dot_LLM-SFT-1.3B	train.csv	FAISS	train.csv	10	프롬프트 500자 제한, epoch 한번에
2024-02-29	0.7056	42dot_LLM-SFT-1.3B	train.csv	FAISS	train.csv	10	프롬프트 500자 제한, epoch 한번에, test split 0.2
2024-03-05	0.6796	42dot_LLM-SFT-1.3B	QA.csv	Ensemble	train.csv	10	증강질문 훈련모델, 앙상블 리트리버
2024-03-05	0.6394	42dot_LLM-SFT-1.3B	QA.csv	FAISS	train.csv	10	증강질문 훈련모델, 페이스 리트리버
2024-03-06	0.7162	42dot_LLM-SFT-1.3B	Q2.csv	Ensemble	train + QA	10	증강 질문2만 사용하여 훈련 (GPT 답변 비중 높음)
2024-03-06	0.7140	42dot_LLM-SFT-1.3B	train + QA	Ensemble	train + QA	6	기존+증강 데이터를 사용하여 훈련
2024-03-07	0.7142	42dot_LLM-SFT-1.3B	train + QA	Ensemble	train + QA	10	기존+증강 데이터를 사용하여 훈련
2024-03-09	0.7284	42dot_LLM-SFT-1.3B	1288 가공	Ensemble	train + QA	100	모델 훈련 케이스 변경_1288개 100epoch
2024-03-10	0.5895	42dot_LLM-SFT-1.3B	1288 다중질문	Ensemble	train + QA	50	50 다중질문 에포크

