

KRX 금융언어모델 경진대회

팀 KRARA

목차

- 1.어떤 데이터를 사용하였는지
- 2.모델 학습 어떻게 하였는지


모델 점수 평가표


	종합	국내기업	재무회계	주가예측	금융에이전트	금융시장	추론시간			
45	kim400/qwen2.5_v7	kim400	0.48	0.68	0.34	0.55	0.47	0.34	624.49	기본 데이터셋 + 생성한 데이터셋 + 스텝 400회, temperature 40
51	ssapi/qwen2.5_v6ssapi		0.47	0.67	0.32	0.54	0.47	0.34	618.64	기본 데이터셋 + 스텝 300회 온도 30
51	ssapi/qwen2.5_v2ssapi		0.47	0.64	0.37	0.54	0.49	0.32	422.18	기본 데이터셋 + 스텝 200회,
54	kim400/krx_qwen2.5_7b_v2	kim400	0.46	0.66	0.32	0.54	0.47	0.32	485.29	기본 데이터셋 + 스텝 60회
65	ssapi/ssap_qwen2.5_v1	ssapi	0.45	0.64	0.34	0.53	0.43	0.32	506.69	기본 데이터셋 + 국내기업 데이터 + 스텝 250회
73	kim400/krx_test	kim400	0.44	0.74	0.27	0.55	0.36	0.3	411.19	라마 + 기본 데이터셋 예폭 60회
89	kim400/krx_qwen2.5_7b_v4	kim400	0.43	0.59	0.28	0.53	0.47	0.29	487.57	기본 데이터셋 + 기본 데이터2 + 기본 데이터셋3 + 180스텝
91	asdkjfhf/qwen7.5b_concatenated		0.43	0.46	0.25	0.55	0.6	0.28	1961.71	
112	kim400/krx_qwen2.5_7b_v3	kim400	0.41	0.54	0.27	0.54	0.45	0.26	594.87	
119	park45er/qwen_v3	park45er	0.4	0.46	0.35	0.5	0.4	0.29	444.08	
121	park45er/qwen2.5_v1	park45er	0.4	0.48	0.26	0.53	0.43	0.29	2059.55	파인튜닝을 한걸 다시 파인튜닝함
267	asdkjfhf/qwen7.5b_concatenated3	asdkjfhf	0.34	0.49	0.25	0.53	0.19	0.24	2710.27	파인튜닝을 한걸 다시 파인튜닝함
276	kim400/qwen_v6	kim400	0.33	0.37	0.27	0.52	0.26	0.25	634.55	직접 만든 데이터셋으로 학습시킴
301	kim400/krx_qwen2-7b-v1	kim400	0.32	0.28	0.23	0.52	0.34	0.23	6662.72	



만든 모델들의 벤치마크 점수를 바탕으로 하여 모델의 개선점을 파악해 나감


기본 베이스 모델은 qwen2.5 7b-instruct를 사용


첫번째 데이터 - 허깅페이스의 amphora/krx-sample-instruction

 **Hugging Face**




 **Datasets:**


 amphora / **krx-sample-instructions**  like 6


Modalities:  Text


Formats:  parquet


Size: 10K - 100K



Libraries:  Datasets  pandas  Croissant

 **Dataset card**



 Viewer

 Files and versions

 Community 2

Dataset Viewer  Auto-converted to Parquet  API

Split (1)
train · 26k rows

prompt string · lengths	response string · lengths
 13 2.88k	 1 3.48k
음악산업의 디지털화가 진행됨에 따라, 전통적인 음반 중심의 수익 모델에서 디지털 음악 서비스 중심으로 변화했음을 설명...	음악산업의 디지털화 주된 원인은 인터넷과 모바일 서비스의 발달로 인해 소비자들이 음악을 접근하고 소비하는 방식이 변...
사후면세점 시장이 최근 몇 년간 급격히 성장한 원인으로는 무엇이 있으며, 이러한 성장세가 지속될 경우 어떤 요소들이 시...	사후면세점 시장이 최근 몇 년간 급격히 성장한 원인으로는 C... 음과 같습니다. 첫째, 해외 관광객의 증가로 인해 면세점의...
휴약값과 산업의 주요 원재료인 후판의 공급량이 특정 국가에 집중되어 있는 상황에서, 이를 해결하기 위한 기업의 전략은 ...	휴약값과 산업의 주요 원재료인 후판의 공급량이 특정 국가에 집중되어 있는 상황에서 기업이 취할 수 있는 전략은 다음과...
정밀 코팅 기술성 신소재는 전자제품, 자동차, 건축 에너지 원...의 분야에서 필수 소재로 여겨지는데, 이러한 신소재의 ...	정밀 코팅 기술성 신소재는 다양한 산업에서 다음과 같은 방...으로 요구 사항을 충족합니다. 1. 전자제품: 전자기에서의...

엔터테인먼트 산업에서 아티스트 매니지먼트가 진행되는 과정에서 아티스트의 상품성뿐 아니라 다양한 유통경로를 개발하여 시장의 부가가치를 창출하기 위해 필요한 요소는 무엇인지, 그리고 이러한 요소들이 아티스트의 성공에 어떻게 기여하는지를 설명하십시오.

아티스트 매니지먼트에서 필요한 요소는 다음과 같습니다. 1. ****브랜드 전략****: 아티스트의 고유한 이미지와 스타일을 확립하여 소비자와의 연결을 강화합니다. 강력한 브랜드는 아티스트의 상품성을 높이고 충성 고객층을 형성하는 데 기여합니다. 2. ****시장 조사****: 대한 소비자 트렌드와 선호도를 파악하여 타겟 마케팅 전략을 수립합니다. 이를 통해 유통경로를 최적화하고 효과적인 진출 전략을 마련할 수 있습니다. 3. ****유통 채널 다양화****: 오프라인과 온라인 플랫폼 등 다양한 유통 경로를 개발하여 아티스트의 콘텐츠가 더 넓은 시장에 도달하도록 합니다. 이렇게 함으로써 판매와 노출 증가를 가져옵니다. 4. ****협업 기회 창출****: 다른 아티스트, 브랜드, 콘텐츠 제작자와의 협업을 통해 시너지 효과를 얻고 새로운 팬층을 확장합니다. 이를 통해 부가가치를 증대시켜 아티스트의 입지를 강화할 수 있습니다. 5. ****팬 관리 및 소통****: 팬과의 적극적인 소통 채널을 유지하여 팬덤을 확장하고 유지합니다. 이는 아티스트의 장기적인 성공에 필수적인 요소입니다. 이러한 요소들이 아티스트의 성공에 기여하는 방식은, 브랜드 인지도를 높여 고객의

2만개가 넘는 증권시장의 대한 주관식 문제가 있기에 이 데이터셋을 기본으로 사용함

해당 데이터에서 추가적인 데이터를 넣어서 학습에 사용

두번째 데이터 - KRX 초보투자자를 위한 증권과 투자



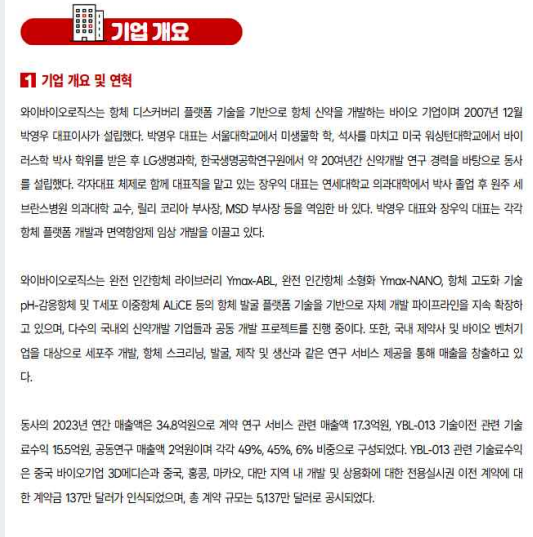
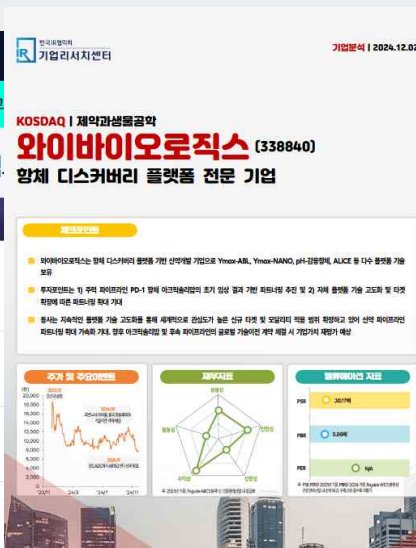
prompt, response

"금융(Finance)이란 무엇인가요?", "금융이란 돈이 남는 사람과 돈이 모자라는 사람 간에 돈(자금)을 돌려쓰는 것을 의미합니다."
"자금의 공급자와 수요자는 각각 누구인가요?", "자금의 공급자는 돈이 남는 사람이고, 자금의 수요자는 돈이 모자라는 사람입니다."
"자금의 흐름은 주로 어떤 경로를 통해 이루어지나요?", "자금은 일반적으로 가계부문(자금의 공급자)으로부터 기업이나 정부(자금의 금융상품의 예시로는 어떤 것들이 있나요?", "금융상품으로는 예금 및 적금, 주식 및 채권과 같은 증권, 선물 및 옵션과 같은 파생상품"
"금융시장이란 무엇인가요?", "금융시장은 금융상품의 거래가 조직적으로 이루어지는 시장으로, 가계, 기업, 정부, 금융기관 등이 자금"
"금융시장의 역할은 무엇인가요?", "금융시장은 경제주체들이 소비나 투자를 원활히 하도록 자금을 조달하고, 자원배분의 효율성을 높"
"금융시장이 경제에 미치는 긍정적인 영향은 무엇인가요?", "금융시장은 경제주체들의 구매력을 향상시키고, 자본형성을 촉진하며, 자"
"금융문맹이란 무엇인가요?", "금융문맹이란 금융 관련 지식 부족으로 돈을 제대로 관리하지 못하여 국민 개인의 삶의 질과 사회 상"
"금융문맹이라는 용어는 언제 등장했나요?", "금융문맹이라는 용어는 1990년대 미국에서 등장했습니다."
"금융문맹의 배경이 된 경제 문제는 무엇이었나요?", "금융문맹의 배경으로 지속적인 저축률 저하, 민간부채 증가, 개인파산 급증 등"
"1990년대 미국에서 금융문맹 문제가 대두된 이유는 무엇인가요?", "당시 미국 경제가 사상 최장기의 고성장을 지속하였으나, 돈의 조"
"금융은 어떤 기준으로 직접금융과 간접금융으로 나눌 수 있나요?", "금융거래가 중개기관을 통해 이루어지느냐에 따라 직접금융과 간"
"직접금융이란 무엇인가요?", "직접금융이란 자금의 수요자가 주식이나 채권을 발행하여 필요한 자금을 직접 조달하는 것을 의미합니"
"간접금융이란 무엇인가요?", "간접금융이란 은행과 같은 금융중개기관이 자금을 조달하고 이를 자금 수요자에게 융통해주는 것을 의"
"금융시장은 어떤 시장들로 구분되나요?", "금융시장은 대출시장, 전통적 금융시장, 외환시장, 상품시장 등으로 구분됩니다."
"대출시장이란 무엇인가요?", "대출시장은 예금취급 금융기관을 통해 자금을 조달하여 최종 자금 수요자에게 공급되는 시장입니다."

KRX에서 제작된 증권과 투자 도서에서 주관식 문제를 제작함

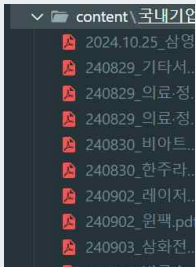
437개의 주관식 문제 생성

세번째 데이터 - KRX 기업 분석 보고서



기업분석 보고서에서 정보를 크롤링함

세번째 데이터 - KRX 기업 분석 보고서



```
def generate_question(text_section):
    prompt = f"""You are a economy specialist, who is very fluent in KOREAN. Generate 1 multiple choice question based on the following text:
    ### 질문: 엑세스바이오의 코로나 진단 제품이 주요 매출에 기여한 이유와, 해당 제품들이 미국 시장에서 어떻게 판매되고 있는지를 설명해 주세
    선택지:
    A. 엑세스바이오의 코로나 진단 제품은 바이러스와 관련 없는 일반적인 건강 검진용으로 개발되었으며, 미국 시장에선 주로 소비자 저가재 방식
    B. 엑세스바이오의 코로나 진단 제품은 대규모 임상시험을 통해 효과가 입증되어, 기존 진단 키트의 80% 이상을 차지하고 있으며, 주로 소매점에
    C. 엑세스바이오의 코로나 진단 제품은 예별라 및 기타 질병 진단에 사용되고 있으며, 미국 시장에서는 주 정부 계약을 통해 독점 판매되고 있음
    D. 엑세스바이오의 코로나 진단 제품은 COVID-19의 급격한 확산에 대응하여 개발된 신속진단키트로, 주로 연한 매출액에서 99.5%를 차지하고 있

    response = client.chat.completions.create(
        model="gpt-4o-mini",
        messages=[{"role": "user", "content": prompt}],
        max_tokens=1000,
        temperature=0.7,
    )
    return response.choices[0].message.content

# Generating answer using ChatGPT API
def generate_answer(text_section, question):
    prompt = f"""Solve this question in Korean: {question} based on the following text:\n{text_section}\n."""

    response = client.chat.completions.create(
        model="gpt-4o-mini",
        messages=[{"role": "user", "content": prompt}],
        max_tokens=1000,
        temperature=0.7,
    )
    return response.choices[0].message.content

qa_pairs=[]
```

Question,Answer

질문: 삼성의 전기차 커패시터용 필름 신규 생산라인이 2025년에 본격 가동될 것으로 예상되는 이유는 무엇인가?

선택지:

- A. 전기차 커패시터용 필름의 수요가 급증하고 있으며, 생산라인이 이미 완공되었기 때문이다.
- B. 글로벌 커패시터용 필름 공급부족 현상으로 인해 새로운 생산라인의 필요성이 대두되고 있으며, 초반막 필름의 판가가 높기 때문이다.
- C. 삼성은 커패시터용 필름 사업에서 적자를 보고 있으며, 신규 생산라인의 가동이 이를 극복할 수 있는 유일한 방법이기 때문이다.
- D. 삼성은 전기차용 필름을 생산하기 위해 해외 업체와 협력하고 있으며, 그 결과로 생산라인이 가동될 예정이다.;### 답변: 질문에 대한 답변은 8입니다.

B. 글로벌 커패시터용 필름 공급부족 현상으로 인해 새로운 생산라인의 필요성이 대두되고 있으며, 초반막 필름의 판가가 높기 때문이다.

이유: 문서에 따르면, 글로벌 커패시터용 필름 공급부족이 지속되고 있으며, 초반막 커패시터용 필름의 가격이 범용 제품을 비해 3~5배 높아 새로운 생산라인의 필요 "다음 문제를 읽고 정답으로 가장 알맞은 것을 고르시오.

질문: 인벤디지랩의 IVL-DrugFluidic 기술은 어떤 주요 특징을 가지고 있으며, 이 기술이 기대되는 이유는 무엇인가?

선택지:

- A. IVL-DrugFluidic 기술은 비만 치료에만 적용되며, 단일 약물의 효과를 극대화하는 데 중점을 두고 있습니다.
- B. IVL-DrugFluidic 기술은 생분해성 고분자와 약물을 조합해 마이크로스피어 형태로 제조하며, 초기 과다방출을 제어하여 안정적인 약물 방출 프로파일을 확보할 수 C. IVL-DrugFluidic 기술은 오직 치매 치료에만 사용될 수 있으며, 다른 질환에는 적용할 수 없습니다.
- D. IVL-DrugFluidic 기술은 마이크로플루이드릭스 기술과 무관하며, 단순히 기존의 주사제 제조 방법을 개선하는 것입니다.;정답은 8입니다.

IVL-DrugFluidic 기술은 생분해성 고분자와 약물을 조합해 마이크로스피어 형태로 제조하며, 초기 과다방출을 제어하여 안정적인 약물 방출 프로파일을 확보할 수 있 "다음 문제를 읽고 정답으로 가장 알맞은 것을 고르시오.

질문: 2025년 반도체와 2차전지 산업의 전망에 대한 주요 내용은 무엇인가?

선택지:

- A. 반도체 산업은 공급 증가로 인해 투자 매력도가 크게 상승할 전망이다.
- B. 2차전지 산업은 전방 수요의 회복으로 기대감이 증가할 것으로 보이며, 반도체 산업은 투자 매력도가 감소할 것으로 전망됩니다.
- C. 반도체 산업은 AI 투자 증가로 인해 수요가 급증할 것으로 예상되며, 2차전지 산업은 성장 둔화가 우려됩니다.
- D. 2차전지 산업은 유럽의 탄소 배출 규제 강화로 인해 급격한 수요 증가가 기대되며, 반도체 산업은 안정적인 성장세를 유지할 것입니다.;2025년 반도체와 2차전지

정답: C. 반도체 산업은 AI 투자 증가로 인해 수요가 급증할 것으로 예상되며, 2차전지 산업은 성장 둔화가 우려됩니다.

이 선택지는 제공된 텍스트의 내용과 일치하며, 반도체 산업의 수요가 AI 투자 증가로 인해 급증할 것으로 예상된다는 점과 2차전지 산업의 성장 둔화에 대한 우려를

질문: 레드캡투어(038390)의 최근 사업 전략 변화에 대해 알맞은 설명은 무엇인가?

선택지:

- A. 레드캡투어는 B2B 중심의 렌터카 사업을 유지하며, 글로벌 시장으로의 확장을 계획하고 있습니다.
- B. COVID-19 팬데믹 이후 소비 트렌드 변화에 따라 렌터카를 선호하는 소비자층이 증가하고 있으며, 구독형 서비스 도입으로 중기적인 성장세를 예상하고 있습니다.
- C. 레드캡투어는 전통적인 여행사로서, 고객의 자동차 소유를 장려하기 위해 마케팅 활동을 강화하고 있습니다.
- D. 레드캡투어는 B2G 사업으로의 확장을 포기하고, 대신 개인 고객을 대상으로 한 서비스를 집중하고 있습니다.

정답: B. COVID-19 팬데믹 이후 소비 트렌드 변화에 따라 렌터카를 선호하는 소비자층이 증가하고 있으며, 구독형 서비스 도입으로 중기적인 성장세를 예상하고 "다음 문제를 읽고 정답으로 가장 알맞은 것을 고르시오.

기업분석 보고서에서 정보를 크롤링함 165개 문항

네번째 데이터 - 인터넷 크롤링 데이터

기업개요

[기준:2024.12.06]

- 1983년 현대전자로 설립, 2001년 하이닉스반도체를 거쳐 2012년 최대주주가 SK텔레콤으로 바뀌면서 SK하이닉스로 상호를 변경함.
- 국내와 중국에 4개의 생산기지와 연구개발법인, 미국, 중국, 홍콩, 대만 등에 판매법인을 운영 중인 인텔의 NAND사업 인수는 1단계 절차를 완료함.
- 2023년 전체 매출 중 DRAM과 NAND가 각각 63%와 29%의 비중을 차지함.

펀더멘털

주요지표	2023/12(A)	2024/12(E)	어닝스트라이프			
			* 단위: 억원, %			
PER	N/A	7.07	재무현월	2024/06	2024/09	
PBR	2.17	1.66	영업이익	컨센서스	51,923.0	67,528.2
PCR	28.74	4.07		잠정치	54,685.4	70,299.6
EV/EBITDA	24.66	3.76		Surprise	5.32	3.95
EPS	-12,517원	23,877원		전년동기대비	289.74	492.31
BPS	77,752원	101,937원	당기순이익	컨센서스	35,279.5	48,056.2
EBITDA	59,433.6억원	362,757.8억원		잠정치	41,202.8	57,486.8
현금DPS	1,200원	1,290원		Surprise	16.79	19.62
현금배당수익률	0.71%	0.76%		전년동기대비	237.75	363.25
회계기준	연결	연결	잠정치 발표(예정)/회계기준	2024/07/25(연결)	2024/10/24(연결)	

(A)는 실적, (E)는 컨센서스

연결 재무제표: 당기순이익 잠정치에서 ○는 전체, ●는 지배주주 기준

```
Unnamed: 0, Symbol, biz_summary, Name, query
```

- 0, A060310, "3S는 1991년 1월 28일에 설립되었으며, 2002년 4월 23일 코스닥시장에 상장함. 반도체 웨이퍼 제조를 주사업으로 하고 있으며, 현재는 반도체 패키징 사업을 집중하고 있음.
- 1, A095576, "AJ네트웍스가 영위하는 사업은 렌탈 사업부문, 창고 및 유통부문, 그리고 기타부문으로 구분됨. 렌탈 사업부문의 매출 비중이 가장 높음.
- 2, A006840, "2012년 9월 민족통일을 통해 제조사업부문을 신설하여 애경유화에 승계하고 지주회사로 전환함.
- 3, A065620, "1994년 9월 반도체 개발, LCD장비, 콘트롤기 및 소프트웨어 제조업을 영위할 목적으로 설립됨.
- 4, A265526, "AP시스템의 주요사업은 장비 연구개발 및 제조 사업으로, AMOLED/LCD 등 디스플레이 장비 제조를 주사업으로 하고 있음.
- 5, A211270, "AP워싱은 2011년 02월 10일에 설립되었으며, 주요 목적사업은 '인공워싱 및 인공워싱 관련 서비스 제공'임.
- 6, A109960, "외피용약, 소화기제, 자양강장변질제 등의 전문의약품을 중심으로 의약품을 유통, 판매하는 회사임.
- 7, A139050, "BF랩스는 2004년 설립되었으며, 메가스터디의 계열회사로 대학생 및 일반성인을 대상으로 한 교육 사업을 영위하고 있음.
- 8, A027410, "BGF는 1990년 편의점 사업을 시작하였으며, 편의점 체인화 사업을 목적으로 1994년 보광, BGF, A282336, "BGF리테일은 편의점 체인화 사업을 영위하고 있으며, BGF리테일의 종속회사는 물류사업, BGF에코머티리얼즈는 1997년 고기능성 폴리머 칩 제조 판매를 목적으로 설립되어, 타 산업 분야로의 사업확장을 추진하고 있음.
- 11, A138930, "BNK금융지주는 2011년 부산은행, BNK투자증권, BNK신용정보, BNK캐피탈이 공동으로 주식출자하여 설립된 회사임.
- 12, A001460, "BYC는 제조업(도소매 포함), 건설업, 임대업을 영위하고 있으며, 제조업은 메리야스류의 의복 제조를 주사업으로 하고 있음.
- 13, A013720, "CBI는 자동차 부품 및 주물제조업 등을 목적으로 1978년 12월 29일에 설립되었음. 자동차 부품 제조를 주사업으로 하고 있음.
- 14, A083790, "2000년 설립되어 2006년 코스닥시장에 상장한 회사로서, 기반기술을 활용하여 신약 후보물질 발굴을 위한 연구를 진행하고 있음.
- 15, A001040, "1953년에 설립되었으며, 2007년 9월 투자사업부문과 제조사업부문을 인적분할하여 지주회사로 전환함.
- 16, A079160, "1999년 3월에 C컨텐츠엔터테인먼트와 호주의 Village Cinema International PTY LTD 2개사와 합작투자하여 설립된 회사임.
- 17, A035760, "1994년 12월 종합유선방송사업자와 홈쇼핑 프로그램의 제작공급 및 도소매업을 목적으로 설립된 회사임.
- 18, A311690, "2009년 설립되어 2019년 코스닥시장에 상장함. 주요 영업사업으로는 마이크로바이옴 치료제 개발을 주사업으로 하고 있음.
- 19, A000120, "1930년 11월에 설립되어 CL사업부문, 택배사업부문, 금융서비스사업부문, 건설사업부문을 중

인터넷 증권 사이트에서 국내기업 정보에 대한 크롤링으로 각 기업에 대한 정보를 추출함 2620개

다섯번째 데이터 - 벤치마크 데이터와 유사한 문제 제작

Question, Answer

"다음 문제를 읽고 정답으로 가장 알맞은 것을 고르시오.

질문: 예코아이의 2024년 실적 전망에 대한 주요 원인은 무엇인가요?

선택지:

- A. 예코아이는 새로운 온실가스 감축 사업으로 인해 매출이 증가할 것으로 예상된다.
 - B. 탄소배출권 국내 판매가격의 하락과 발급 지연이 실적 부진의 주요 원인으로 작용하고 있다.
 - C. 예코아이의 환경친설팅 사업이 매출의 대부분을 차지하고 있다.
 - D. 국제 기후변화 체제의 재정립이 완료되어 실적 회복이 가능할 것으로 보인다."
- "문제에서 요구하는 예코아이의 2024년 실적 전망에 대한 주요 원인은 다음과 같습니다.

주어진 텍스트에 따르면, 2024년 실적 부진의 주요 원인은 ""탄소배출권 국내 판매가격의 하락과 탄소배출권 발급 지연""입니다. 따라서 가장 알맞은 선택지는 다음과 같습니다:

B. 탄소배출권 국내 판매가격의 하락과 발급 지연이 실적 부진의 주요 원인으로 작용하고 있다."

"### 질문: 그린리소스(402490)의 초고밀도(SD) 특수코팅 기술이 반도체 시장에서 증가하는 수요에 어떻게 기여하고 있는지를 설명해 주세요.

선택지:

- A. 그린리소스의 SD 특수코팅 기술은 5nm 이상의 공정에서 사용되며, 식각 공정 환경이 개선되어 수확성이 증가하고 있습니다.
 - B. 그린리소스의 SD 특수코팅 기술은 반도체 제조 과정에서 필요 없는 부품으로, 수요가 감소하고 있습니다.
 - C. 그린리소스의 SD 특수코팅 기술은 5nm 이하의 선단 공정에서 사용되며, EUV 및 High-K 사용 증가로 인해 식각 공정 환경이 가혹해짐에 따라 수요가 증가하고 있습니다.
 - D. 그린리소스의 SD 특수코팅 기술은 단순한 보호코팅으로, 반도체 시장에서는 별다른 영향을 미치지 않고 있습니다."
- "그린리소스(402490)의 초고밀도(SD) 특수코팅 기술이 반도체 시장은 다음과 같습니다.

답변:

- C. 그린리소스의 SD 특수코팅 기술은 5nm 이하의 선단 공정에서 사용되며, EUV 및 High-K 사용 증가로 인해 식각 공정 환경이 가혹해짐에 따라 수요가 증가하고 있습니다.

이유:

- 그린리소스의 SD 코팅 기술은 5nm 이하의 초미세 공정에 적합하며, 이러한 공정에서는 EUV(극자외선) 및 High-K 물질의 사용이 증가하여 식각 공정 환경이 더욱 가혹해집니다.
- 이로 인해 반도체 제조에서 필요한 핵심 부품에 대한 초고밀도 특수코팅의 수요가 증가하고 있습니다.
- 보고서에 따르면, 3nm 이하의 초미세 공정 채택 증가에 따라 SD코팅의 수요도 급증할 것으로 전망되고 있습니다."

"### 질문: 헥토이노베이션의 2024년 연결 실적 전망에 따르면, 매출액과 영업이익은 각각 얼마로 예상되고 있습니까?

선택지:

- A. 매출액 3,110억원, 영업이익 445억원

예시 벤치마크 문제와 유사한 문제를 제작함

객관식으로 데이터 제작, 답변 제작시 풀이에 대한 이유도 같이 제작

	종합	국내기업	재무회계	주가예측	금융에이전트	금융시장	추론시간			
45	kim400/qwen2.5_v7	kim400	0.48	0.68	0.34	0.55	0.47	0.34	624.49	기본 데이터셋 + 생성한 데이터셋 + 스텝 400회, temperature 40
51	ssapi/qwen2.5_v6	ssapi	0.47	0.67	0.32	0.54	0.47	0.34	618.64	기본 데이터셋 + 스텝 300회 온도 30
51	ssapi/qwen2.5_v2	ssapi	0.47	0.64	0.37	0.54	0.49	0.32	422.18	기본 데이터셋 + 스텝 200회,
54	kim400/krx_qwen2.5_7b_v2	kim400	0.46	0.66	0.32	0.54	0.47	0.32	485.29	기본 데이터셋 + 스텝 60회
65	ssapi/ssap_qwen2.5_v1	ssapi	0.45	0.64	0.34	0.53	0.43	0.32	506.69	기본 데이터셋 + 국내기업 데이터 + 스텝 250회
73	kim400/krx_test	kim400	0.44	0.74	0.27	0.55	0.36	0.3	411.19	라마 + 기본 데이터셋 에폭 60회
89	kim400/krx_qwen2.5_7b_v4	kim400	0.43	0.59	0.28	0.53	0.47	0.29	487.57	기본 데이터셋 + 기본 데이터2 + 기본 데이터셋3 + 180스텝
91	asdkjfh/qwen7.5b_concatenated		0.43	0.46	0.25	0.55	0.6	0.28	1961.71	
112	kim400/krx_qwen2.5_7b_v3	kim400	0.41	0.54	0.27	0.54	0.45	0.26	594.87	
119	park45er/qwen_v3	park45er	0.4	0.46	0.35	0.5	0.4	0.29	444.08	
121	park45er/qwen2.5_v1	park45er	0.4	0.48	0.26	0.53	0.43	0.29	2059.55	파인튜닝을 한걸 다시 파인튜닝함
267	asdkjfh/qwen7.5b_concatenated3	asdkjfh	0.34	0.49	0.25	0.53	0.19	0.24	2710.27	파인튜닝을 한걸 다시 파인튜닝함
276	kim400/qwen_v6	kim400	0.33	0.37	0.27	0.52	0.26	0.25	634.55	직접 만든 데이터셋으로 학습시킴
301	kim400/krx_qwen2-7b-v1	kim400	0.32	0.28	0.23	0.52	0.34	0.23	6662.72	

앞선 데이터셋들을 이용하여 데이터셋을 바꿔가며 데이터셋 제작



그 결과 amphora 데이터셋에 KRX 금융 자료와 기업분석보고서를 합친 데이터가 가장 높은 점수를 냈다

krara

33.07

45.6

53.6

7B

0

앞에서 넣은 모든 데이터를 넣었는데 오히려 결과가 좋지 않았다

이때 느끼는

1.비슷한 내용이 반복되는 데이터다 보니 오히려 모델에 악영향을 주었을 것 같다

2.amphora 데이터의 양이 너무 많았기에 비율이 맞지 않아 학습이 안 된 것 같다

["prompt, response"]

""금융(Finance)이란 무엇인가요?, ""금융이란 돈

""자금의 공급자와 수요자는 각각 누구인가요?, ""

""자금의 흐름은 주로 어떤 경로를 통해 이루어지

""금융상품의 예시로는 어떤 것들이 있나요?, ""금

""금융시장이란 무엇인가요?, ""금융시장은 금융

""금융시장의 역할은 무엇인가요?, ""금융시장은

""금융시장이 경제에 미치는 긍정적인 영향은 무

""금융문맹이란 무엇인가요?, ""금융문맹이란 금융

""금융문맹이라는 용어는 언제 등장했나요?, ""금

""금융문맹의 배경이 된 경제 문제는 무엇이었나

""1990년대 미국에서 금융문맹 문제가 대두된 이

""금융은 어떤 기준으로 직접금융과 간접금융으로

""직접금융이란 무엇인가요?, ""직접금융이란 자

""간접금융이란 무엇인가요?, ""간접금융이란 은

모델 학습 과정

1. qwen2.5 7b-it에 스텝을 400회로 사용, warmup_steps은 40으로 사용
2. 에폭을 1,3회씩 학습한 모델이 오히려 성능이 좋지 않았다

```
from unsloth import is_float16_supported

trainer = SFTTrainer(
    model = model,
    tokenizer = tokenizer,
    train_dataset = dataset,
    dataset_text_field = "text",
    max_seq_length = max_seq_length,
    dataset_num_proc = 2,
    packing = False, # Can make training 5x faster for short sequences.
    args = TrainingArguments(
        per_device_train_batch_size = 2,
        gradient_accumulation_steps = 4,
        warmup_steps = 40,
        # num_train_epochs = 1, # Set this for 1 full training run.
        num_train_epochs = 400
```

모델 학습과정에서 알게 된 사실

9 444.08
9 2059.55 파인튜닝을 한걸 다시 파인튜닝함

```
[ ] if False:
    # I highly do NOT suggest - use Unsloth if possible
    from peft import AutoPeftModelForCausalLM
    from transformers import AutoTokenizer
    model = AutoPeftModelForCausalLM.from_pretrained(
        "lora_model", # YOUR MODEL YOU USED FOR TRAINING
        load_in_4bit = load_in_4bit,
    )
    tokenizer = AutoTokenizer.from_pretrained("lora_model")
```

파인튜닝을 한 모델을 한번 더 파인튜닝을 해보았는데 추론 시간이 비이상적으로 늘어남
이는 lora를 이용해서 파인튜닝을 한 모델을 한번 더 lora를 이용하면
이전에 학습한 내용이 변형된다

결론

1. 데이터셋을 제작하는 과정에서 데이터를 더 많이 넣는다 하여 좋은 결과가 나오지 않음
2. 오히려 일관된 형태의 데이터는 모델의 학습에 방해가 됨
3. epoch으로 학습을 하더라도 성능이 더 잘 나오지 않았다, 오히려 step으로 할 경우 성능이 더 좋았다
4. 특정 형식의 데이터셋 양이 너무 적다면 학습에 영향을 줄 수 없기에 다양한 데이터의 비율을 맞춰가야함

감사합니다