

**SK네트웍스 Family AI과정 14기  
 모델링 및 평가 수집된 데이터 및 전처리 문서**



**□ 개요**

* 산출물 단계 : 모델링 및 평가
* 평가 산출물 : 수집된 데이터 및 전처리 문서
* 제출 일자 : 2025.09.12
* 깃허브 경로 : https://github.com/skn-ai14-250409/SKN14-Final-3Team-AI
* 작성 팀원 : 이승혁

## 1. 개요

### 1.1 데이터셋 개요

본 프로젝트는 KB금융 특화 sLLM(smaller Large Language Model)을 파인튜닝하기 위해 구축된 데이터셋에 대해 기술합니다. 이 데이터셋은 모델이 금융 도메인 지식, 특히 KB금융그룹의 페르소나를 효과적으로 학습하도록 설계되었습니다.

이 섹션에서는 데이터셋의 구성 요소, 구축 목표, 그리고 핵심 특징에 대해 설명합니다. 이어지는 장에서는 데이터의 구체적인 출처, 수집 방법, 전처리 파이프라인, 그리고 품질 관리 및 저장 방식에 대해 상세히 다룰 것입니다.

### 1.2 데이터 수집 목적

본 파인튜닝의 목적은 **Qwen2.5-14B-Instruct** 모델을 특정 도메인(한국 금융)과 페르소나(KB금융 어시스턴트)에 맞게 최적화하는 것입니다. 이를 위해 다음과 같은 능력을 모델에 주입하는 것을 목표로 데이터를 수집 및 가공하였습니다.

* **금융 전문성:** 금융 용어, 시장 분석, 상품 정보 등 전문적인 질의에 대해 정확한 답변을 생성하는 능력.
* **자연스러운 대화 능력:** 딱딱한 정보 전달을 넘어, 사용자와 자연스럽게 상호작용하는 능력.
* **KB 페르소나 체화:** KB금융그룹의 실제 문서와 데이터를 학습하여, 신뢰할 수 있고 일관된 KB만의 목소리로 소통하는 능력.

## 2. 데이터 수집

### 2.1 데이터 출처

### 모델의 전문성과 범용성을 동시에 확보하기 위해, 크게 세 가지 유형의 데이터 소스를 활용하였습니다.

### 금융 도메인 데이터 (금융계 핵심 지식):

### 출처: aiqwe/FinShibainu (KRX LLM 경진대회 우수작 학습 데이터)

### 내용: 금융 Q&A, 시장 분석 등 전문 지식.

### 

### 일반 대화 데이터 (자연스러움 및 품질):

### 출처: beomi/KoAlpaca-RealQA

### 내용: 모델의 페르소나 왜곡을 방지하고 다양한 질문 형태를 학습시키기 위해, 품질이 검증된 KoAlpaca 데이터셋을 사용.

### 

### KB 특화 데이터 (페르소나 및 도메인 지식):

### 출처: sumilee/SKN14-Final-3Team-Data2 (KB금융그룹 공개 문서 233건)

### 내용: 강령, 법률, 상품, 약관 등 실제 업무 문서.

### 2.2 데이터 유형

* 수집된 모든 데이터는 텍스트 형식이며, 파인튜닝을 위해 대화형 Q&A 구조로 가공되었습니다

#### 2.3 데이터 규모

## 데이터 조합 전략: 멘토의 조언에 따라, 모델의 대화 능력과 도메인 전문성의 균형을 맞추기 위해 새로운 비율로 데이터를 구성하였습니다.

## 일반 대화 데이터: 15,000건 샘플링

## 금융 도메인 데이터: 15,000건 샘플링

## KB 특화 데이터: KB 문서 233건을 OpenAI API(gpt-4o-mini)를 활용하여 1,157건의 Q&A 데이터셋으로 자동 생성 후, 전량 사용.

## 최종 데이터셋: 위 데이터를 통합하여, 총 31,157건의 데이터를 확보.

## 최종 분할: 훈련용 29,599건 (95%), 평가용 1,558건 (5%).

## 허깅페이스 링크: [rucipheryn/combined-dataset-30K-final-v2](https://huggingface.co/datasets/rucipheryn/combined-dataset-30K-final-v2)

## 3. 데이터 전처리

#### 3.1 전처리 개요

* **데이터 정제:** 다양한 소스에서 수집한 데이터의 구조를 표준화.
* **파인튜닝 형식 변환:** 모든 데이터를 **의 공식 토크나이저**를 사용하여 Qwen 챗 템플릿에 맞는 단일 텍스트 형식으로 변환.
* **데이터 통합 및 분할:** 최종 데이터를 훈련용과 평가용으로 분할하여 데이터 유출 방지.

#### 3.2 데이터 정제 (Format Unification)

* **목표:** 서로 다른 구조를 가진 3개의 데이터셋을 [{"role": "user", ...}, {"role": "assistant", ...}] 형식의 단일 대화 구조로 통일.
* **구현:** 각 데이터셋의 질문/답변 컬럼(question/answer\_B 또는 question/answer)을 user/assistant 역할로 매핑하는 커스텀 함수를 datasets.map에 적용.
* **결과:** 각 소스별 데이터가 후속 처리가 용이한 일관된 대화 리스트 형식으로 정제됨.

#### 3.3 파인튜닝 형식 변환 (Chat Template Application)

* **목표:** 통일된 대화 형식 데이터를, 모델이 실제 학습에 사용할 최종 텍스트 문자열로 변환.
* **구현:**
  + **Qwen/Qwen2.5-14B-Instruct의 공식 AutoTokenizer**를 로드.
  + tokenizer.apply\_chat\_template 함수를 사용하여, 모든 대화 데이터를 <|im\_start|>user...<|im\_end|>\n<|im\_start|>assistant...<|im\_end|> 형태의 Qwen 공식 템플릿으로 변환. 이 과정을 통해 **EOS 토큰 불일치 문제를 원천적으로 해결**.
* **결과:** 모든 데이터가 단일 text 컬럼을 가진, 파인튜닝에 즉시 사용 가능한 형태로 가공됨.

#### 3.4 학습, 평가 데이터 분할

* **목표:** 여러 소스에서 정제된 데이터를 하나의 최종 데이터셋으로 합치고, 훈련용과 평가용으로 분리.
* **구현:**
  1. **통합:** datasets.concatenate\_datasets 함수를 사용하여 준비된 3종의 데이터셋 (총 31,157건)을 통합.
  2. **셔플링:** final\_dataset.shuffle(seed=42)를 통해 데이터 순서에 대한 편향을 방지.
  3. **분할:** train\_test\_split(test\_size=0.05) 함수를 사용하여 95:5 비율로 데이터 분할.
* **결과:** 훈련셋(29,599건)과 평가셋(1,558건) 확보.

#### 3.5 데이터 저장

* **목표:** 재현 가능하고 팀원 간 공유가 용이하도록, 최종 데이터셋을 허깅페이스 허브에 저장.
* **구현:** final\_dataset\_dict.push\_to\_hub()를 사용하여, 분할된 train 및 test 스플릿을 지정된 저장소에 게시.
* **결과:** **rucipheryn/combined-dataset-30K-final-v2** 에 최종 데이터셋이 성공적으로 업로드됨.

### 4. 향후 사용 계획

본 프로젝트를 통해 성공적으로 구축된 KB금융 특화 sLLM과 RAG 평가 파이프라인을 기반으로, 다음과 같은 단계별 고도화 및 확장 계획을 수립하였습니다.

#### 4.1 모델 성능 고도화 (Model Enhancement)

* **하이퍼파라미터 최적화:** 현재 구축된 Baseline 모델을 기준으로, 학습률(Learning Rate), 배치 사이즈(Batch Size), 훈련 에포크(Epochs) 등 핵심 하이퍼파라미터에 대한 체계적인 실험을 진행하여 최적의 조합을 탐색할 계획입니다. 모든 실험 과정과 결과는 **WanDB**를 통해 기록하고 정량적으로 비교 분석할 것입니다.
* **추가 사전 훈련 (Continued Pre-training):** 모델의 근본적인 금융 도메인 이해력을 강화하기 위해, "무작위 단어 마스킹(MLM)" 기법을 활용한 추가 사전 훈련(CPT) 및 고객언어 가이드를 활용하여 DPO까지 진행할 예정입니다. 또한 금융 뉴스, 증권사 리포트 등 대량의 비정형 텍스트 데이터를 학습시켜, 모델이 최신 금융 용어와 미묘한 시장의 뉘앙스까지 이해하도록 성능을 개선할 계획입니다.

#### 4.2 RAG 시스템 연동 및 평가

* **실시간 RAG 파이프라인 구축:** 현재 파인튜닝된 모델을 검색 증강 생성(RAG) 시스템과 연동하여, KB금융그룹의 방대한 내부 문서(PDF)를 실시간으로 참조하여 답변을 생성하는 프로토타입을 개발할 계획입니다. 이를 위해 Upstage Document AI, OpenAI batch api와 같은 고성능 문서 파서와 Q&A셋 제작 도구를 활용하여 벡터 DB를 구축할 것입니다.
* **RAG 성능 정량 평가:** Ragas 프레임워크를 사용하여, 구축된 RAG 시스템의 성능을 Faithfulness(충실성), Answer Relevancy(답변 관련성) 등 객관적인 지표로 측정하고 지속적으로 개선해 나갈 것입니다.

#### 4.3 데이터셋의 지속적인 관리 및 확장

* **주기적인 데이터 업데이트:** 금융 시장의 최신 트렌드와 새로운 금융 상품 정보를 모델이 반영할 수 있도록, 분기별 또는 반기별로 신규 데이터를 수집하고 파인튜닝 데이터셋을 업데이트하는 프로세스를 정립할 계획입니다.
* **데이터 품질 관리:** LLM-as-a-Judge 기법을 도입하여, 새로 추가되는 데이터의 품질을 반자동으로 검수하고, 일관성 있는 고품질 데이터셋을 유지하기 위한 시스템을 구축할 예정입니다.