그래픽, 폰트, 그래픽 디자인, 로고이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

| **프로젝트 주제** | 자체 sLLM을 활용한 은행 사내 업무 AI 플랫폼 | | |
| --- | --- | --- | --- |
| **데이터 출처** | KB 국민은행, 여신금융협회, 금융감독원, Kaggle, huggingface | **데이터 형태** | PDF 파일, CSV 파일 |
| **수집데이터 개요** | KB 국민은행의 여신 사업을 진행하게 될 은행 사내 직원이 여신 심사 업무 지원 플랫폼을 통해 금융 관련 질문과 답변(QnA)을 수행하고, 고객 심사 및 문서 기반 분석·보고서를 생성할 수 있는 플랫폼을 구성하기 위해 데이터를 수집함.  챗봇(자체 sLLM)이 정확한 답변 및 출처를 제공하기 위해 최신 법령/제도의 문서(PDF)위주 데이터 수집과 KB 국민은행의 내규 데이터와 여신업무 관련 문서 위주로 수집하였으며, 머신러닝 기반 대출 적정성 평가와 심사 결과 보고서 작성을 하기 위해 학습에 필요한 ‘대출 승인 분류 데이터 세트’를 Kaggle에서 수집함. | | |
| **데이터 출처** | KB 금융 블로그 <https://kbthink.com/loan.html> KB 국민은행 고객센터 <https://obank.kbstar.com/quics?page=C019999&cc=b034360:b034261> KB 금융그룹 웹사이트 <https://www.kbfg.com/kor/ir/report/annual/list.jsp>  여신금융협회 <https://www.crefia.or.kr/portal/infocenter/regulation/selfRegulation.xx>  금융감독원  <https://www.fss.or.kr/fss/bbs/B0000111/list.do?menuNo=200139>  금융위원회  <https://www.fsc.go.kr/index>  Kaggle - 개인 신용 대출(Loan Apporoval Classification) <https://www.kaggle.com/datasets/taweilo/loan-approval-classification-data/data>  Kaggle - 주택 담보 대출 (Home Loan Approval) <https://www.kaggle.com/datasets/rishikeshkonapure/home-loan-approval>  금융 QA/MCQA데이터  <https://huggingface.co/datasets/aiqwe/FinShibainu> | | |
| **원본 데이터 예시** | 텍스트, 스크린샷, 폰트, 웹 페이지이(가) 표시된 사진  AI가 생성한 콘텐츠는 부정확할 수 있습니다.텍스트, 스크린샷, 평행, 번호이(가) 표시된 사진  AI가 생성한 콘텐츠는 부정확할 수 있습니다.  **<KB 금융 블로그 자주 묻는 질문> <금융상품설명서 PDF >** | | |
| **데이터 전처리 시 유의할 점** | 데이터 전처리 과정에서는 PDF 문서를 단순히 불러오는 것만으로는 충분하지 않다는 점을 유의해야 함. 예를 들어, PDF 문서의 특성상 표, 이미지, 각주, 머리말·꼬리말, 불필요한 공백이나 줄바꿈 등이 그대로 텍스트로 추출되면 문장이 끊기거나 의미가 왜곡될 수 있음. 또한 약관, 규정, 상품설명서와 같은 금융 문서들은 법률적 용어와 긴 문장이 많아 작은 오류만 있어도 검색이나 질의응답 시 부정확한 결과가 발생할 수 있음. 이러한 문제를 해결하기 위해서는 단순 로더(loader)를 통한 불러오기보다는 파서(parser)를 활용하여 텍스트 구조를 보다 정교하게 추출하고, 이후 불필요한 기호 제거, 문장 단위 분할, 표준화와 같은 전처리를 거쳐야 함. 최종적으로는 이러한 전처리 과정을 통해 데이터의 품질을 확보함으로써 RAG 시스템에서의 검색 정확도를 높이고, 파인튜닝 데이터셋 역시 안정적으로 구축할 수 있음.  또한 sLLM 학습 데이터 생성 시, 선택한 모델(Qwen2.5-7B-Instruct)에 맞는 토크나이저와 스페셜 토큰을 적용하여 모델 전용 데이터셋으로 제작해야 함. 또한 Q&A셋으로 만들어 질문에 대한 답변을 준비할 수 있도록 해야 함. | | |