SK네트웍스 **Family AI**과정 **3**기

모델링 및 평가 수집된 데이터 및 전처리 문서

☐ 개요

- 산출물 단계 : 모델링 및 평가
- 평가 산출물 : 수집된 데이터 및 전처리 문서
- 제출 일자 : 12월 17일
- 깃허브 경로 : https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN03-FINAL-3Team
- 작성 팀원 : 송영빈, 장수연, 오승민

개요	**1.** 목적 모델링 및 평가 단계에서 데이터의 품질을 확보하고, 데이터가 학습 및 평가에 효과적으로 사용될 수 있도록 하기 위함 **2.** 특징 ● 기존 전처리 과정에서 얻어진 데이터를 바탕으로 모델 학습에 적합한 형태로 추가 정제 ● 데이터 검증을 통해 모델링 및 평가에 필요한 신뢰성과 품질 확보
데이터 검증	**1.** 검증 기준 ● 데이터의 정확성: FAQ 데이터와 원본 규정 문서 간의 일치 여부 검토 ● 일관성: 동일한 질문 카테고리에 대해 유사한 응답이 생성되었는지 확인 ● 중복 제거: 중복 질문-응답 페어 제거 ● 결측지 처리: 필수 필드의 누락 데이터 제거 **2.** 검증 프로세스 ● 전처리된 데이터셋의 샘플링을 통해 데이터 품질 확인 ● 모델 평가 중 환각(hallucination) 응답 여부를 판별 ● 필요시 수작업으로 수정 및 보완

데이터 저장 및 관리	

데이터 전처리 과정	**1. PDF** 데이터 전처리 ● PDF에서 텍스트를 추출하고 장/조/주제/내용별로 구조화 ● 불필요한 공백 및 특수문자 제거필요한 공백 및 특수문자 제거 ● 구조화 된 텍스트를 JSONL 형태로 저장 **2. FAQ** 데이터 생성 및 전처리 ● **1**차 **FAQ** 데이터 생성 ○ PDF 텍스트 추출 및 항목 별로 분할 ○ LLM을 활용하여 분할된 항복 별로 다양한 표현의 질문과 답변 한번에 생성 ○ 1차 생성 문제점: 질문과 답변을 같은 파라미터를 통해 생성하여 환각 응답의 분포가 많음 ● **2**차 **FAQ** 데이터 생성 ○ 질문과 답변을 각각 다른 파라미터를 사용하여 따로 생성 ○ 환각 응답의 분포가 감소 ○ 문제점: 답변이 문서 전체를 참조하지 않고 추출된 항목에 대해서만 참조하여 답변을 생성해서 문서에 해당하는 내용이 있지만 답변할 수 없다고 하는 경우 발생 ● **3**차 **FAQ** 데이터 생성 ○ 기존에 생성된 질문에 대한 답변을 FAISS를 이용해 임베딩된 규정 데이터에서 관련 규정을 검색 ○ 검색된 정보를 기반으로 LLM을 통해 답변 생성 ● 데이터 검증 ○ 환각 응답 여부 및 정확성 검토 ○ 잘못된 답변에 대하여 수작업 수정 **3. HR** 데이터 전처리 ● Python과 Faker 라이브러리를 활용해 근태, 복지 포인트 등 시뮬레이션 데이터 생성

	● 직급, 부서, 팀 등 텍스트로 사용되던 데이터 코드화하여 관리하도록 변경 ● 코드화를 통해 데이터 일관성 및 확장성 확보
데이터 전처리 결과	**PDF** 데이터: ● 문서 수: 4개 ● 항목(장, 조, 주제) 수: 192개 ● 데이터 형식: JSONL (장/조/주제/내용) **FAQ** 데이터: ● 총 질문-응답 쌍: 약 1600개 ● 데이터 분리: 학습용(Train) 1300개, 검증용(Validation) 300개 ● 데이터 품질 향상을 위한 검증 및 수정 완료 **HR** 데이터: ● 가상 직원 수: 약 700명 ● 데이터 구성: 근태 관리, 복지 포인트, 개인화된 답변 생성에 활용 가능한 데이터셋