

**SK네트웍스 Family AI과정 15기  
 데이터 수집 및 저장 수집 데이터 보고서**



| **산출물 단계** | 데이터 수집 및 저장 |
| --- | --- |
| **평가 산출물** | 수집 데이터 보고서 |
| **제출 일자** |  |
| **깃허브 경로** | 깃허브 주소 기재 |
| **작성 팀원** |  |

1. **수집 데이터 개요**

| **데이터명** | **수집 대상** | **수집 목적** | **사용 예정 기능** | **출처/저작권** |
| --- | --- | --- | --- | --- |
| 예: 상담 로그 샘플 | LLM 챗봇 사용자 | 프롬프트 튜닝, 모델 응답 테스트 | 프롬프트 리파인먼트 | 자체 생성, 내부 테스트 |

1. **수집 방법 및 자동화 절차**

* 수집 방식 (해당 항목에 체크)  
  + 웹 크롤링
  + API 호출
  + 사용자 입력
  + 문서 파일 업로드 (PDF, CSV 등)
  + 기타: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_
* 수집 도구 또는 스크립트 설명:  
  + 사용한 언어/라이브러리: (예: Python + BeautifulSoup, requests 등)
  + 자동화 여부 및 주기:
  + 오류 발생 시 예외 처리 전략:
* 예시 스크립트 또는 흐름도 첨부: (이미지, 순서도 또는 코드)

1. **데이터 설명 및 구성**

**4.1 파일 및 필드 설명**

| **파일명 또는 테이블명** | **필드명** | **데이터 타입** | **설명** | **예시** |
| --- | --- | --- | --- | --- |
| chat\_logs.csv | prompt\_text | string | 사용자 질문 | “오늘 날씨 어때?” |
|  | timestamp | datetime | 수집 시간 | 2025-05-10 15:01:22 |

**4.2 데이터 양**

* 전체 수집 데이터 건수:
* 추출된 고품질 데이터 건수 (필터링 후 기준):

**4.3 저장 위치 및 포맷**

* 저장 경로: (예: /data/raw/chat\_logs/)
* 저장 포맷: CSV / JSON / SQLite / 기타: \_\_\_\_\_\_\_\_\_\_
* 인코딩: UTF-8 / 기타

1. 법적·윤리적 검토

* 개인정보 포함 여부:  
  + 포함  [ ] 미포함
  + 포함된 경우 필드: (예: 이름, 전화번호 등)
* 비식별화 조치 여부:  
  + 이름 마스킹  [ ] 고유키 대체  [ ] 기타
* 출처 및 사용권:  
  + 공개 여부: [ ] 공개  [ ] 내부사용 한정
  + 라이선스 또는 약관 검토 여부:
* 검토자 및 검토 일자:

**5. 데이터 품질 및 정합성 관리 방안**

* 중복 제거 기준: (예: user\_id + timestamp 기준 중복 제거)
* 정합성 검증 방법:  
  + 예: 잘못된 날짜 형식 필터링, 필수 필드 누락 제거
* Null 처리 및 결측치 전략:
* 표준화 전략:  
  + 텍스트 전처리 여부 (예: 소문자 변환, 특수문자 제거 등)

**6. 변경 이력 및 보완 내역**

| **변경일** | **변경자** | **변경 내용** | **비고** |
| --- | --- | --- | --- |
| 2025-05-09 | 홍길동 | 전화번호 필드 삭제 및 익명화 처리 | 개인정보 보호 강화 조치 |