SK네트웍스 Family AI 과정 20기

데이터 전처리 인공지능 학습 결과서

산출물 단계	데이터 전처리
평가 산출물	인공지능 데이터 전처리 결과서
제출 일자	2026년 2월 2일
깃허브 경로	SKN20-FINAL-5TEAM
작성 팀원	김황현, 최소영

1. 문서 개요

- **프로젝트명**: Engineering Gym
- **전처리 목적**:
  - 전처리 과정의 재현성과 이해도를 높이고, LLM 생성 데이터를 시스템 규격에 맞게 표준화함
  - 플랫폼 내 축적된 사용자의 시스템 구성(Architecture) 데이터와 파이썬 코드 실행 로그를 정제하여, 사용자의 설계 의도를 파악하고 문제 해결 패턴을 분류하는 분류(Classification) 및 추천 모델의 학습 데이터 품질을 확보함
- **문제 정의**:
  - 외부 기술 블로그 및 영상 기반으로 AI가 생성한 비정형 데이터의 논리적 오류를 검증하고 구조화함
  - 비정형 코드 데이터: 사용자가 입력한 파이썬 코드 및 슈도코드의 노이즈(주석, 들여쓰기 불일치) 제거
  - 그래프 기반 데이터: Mermaid.js 포맷이나 JSON 형태의 시스템 아키텍처 연결 정보를 AI 모델이 이해 가능한 인접 행렬(Adjacency Matrix) 또는 피처 벡터로 변환하기 위한 정규화

2. 데이터셋 개요

2.1 데이터 수집 경로 및 참고 문헌

● 실무 현장감을 반영하기 위해 검증된 기술 매체를 참고하여 문제를 설계하고 수집하였습니다.

- 기술 블로그: 네이버, 카카오, 구글, 우아한 형제들, 당근 등 주요 기업의 실제 디버깅 사례 참고
- 학습 플랫폼: NeetCode의 문제 구조 및 허깅페이스(Hugging Face) 데이터셋 파이프라인 참고
- 영상 콘텐츠: 구글/메타 시스템 아키텍처 면접 영상 및 유튜브 의사코드 면접 영상 활용

2.2 데이터 구성 및 항목 (JSON 구조화 방식)

● 수집된 비정형 정보를 시스템 표준 스키마에 맞게 다음과 같이 JSON 형식으로 구조화하여 저장합니다.

● 시스템 운영에 핵심적인 의사코드(Logic Mirror)와 디버깅(Bug Detective), 시스템 아키텍처(System Architecture) 데이터를 중심으로 구성됩니다.

항목명	설명	예시
id	문제 고유 식별자 (UUID)	550e8400-e29b-41d4-a716-446655440000
level	level 문제 난이도 (1~5단계)	3
title	문제 제목	"Instagram Home Feed"
logic_type	논리 유형 (순차, 선택, 반복 등)	SEQUENCE, SELECTION
scenario	문제 상황 설명 (Context)	"미래 데이터가 학습에 포함되지 않도록..."

원본 데이터 샘플 (예시 5건)

실제 progressive-problem.json 및 stage.js 파일의 구성을 바탕으로 한 샘플입니다.

1. **ID 101**: [Level 1] 씨앗이 꽃이 되는 과정 (Logic: SEQUENCE)
2. **ID 102**: [Level 2] 조건문에 따른 이동 경로 찾기 (Logic: SELECTION)
3. **ID 201**: [Level 3] 스포티파이 시스템 아키텍처 설계 문제
4. **ID 301**: [Level 4] 데이터 전처리 파이프라인 버그 수정 (Scenario: Data Leakage)
5. **ID 302**: [Level 5] 대규모 트래픽 처리를 위한 서버 부하 분산 설정

3. 전처리 프로세스 개요

- **전체 흐름도**: ① 외부 데이터 수집(LLM 생성) → ② 구조 표준화(JSON 스키마 검증) → ③ 결측치 처리 → ④ 정규화 및 보안 암호화 → ⑤ 무결성 검증
- **구조화 방식**: Claude 3.5 Sonnet 모델을 활용하여 비정형 텍스트를 scenario, correct_code, hint 등 정해진 필드로 분리하여 저장

4. 세부 전처리 단계

4.1 결측치 및 이상치 처리

- 결측치 처리: id, title, level 등 필수 필드가 누락된 데이터는 시스템 안정성을 위해 제거함
- 이상치 처리: 난이도 범위를 벗어난 데이터 보정 및 전문가 리뷰를 통한 정합성 검토

4.2 정규화 및 표준화 결과

- 보안 표준화: 사용자 비밀번호에 PBKDF2 알고리즘 적용 및 평문 저장 금지 정책 수립
- 구조 정규화: 모든 테이블에 use_yn(Soft Delete), create_date 등 공통 감사 필드 적용
- 텍스트 정제: LLM 생성 텍스트의 불필요한 특수문자 및 이스케이프 문자 제거

4.3 데이터 변환 및 생성

- 레이블 인코딩: JOB_ROLE(직군) 및 LOGIC_TYPE(논리 유형)을 공통 코드로 변환하여 관리합니다.
- 파생 변수: 학습 동기 부여를 위해 total_points(누적 포인트) 및 streak_days(연속 활동일) 필드를 생성하여 관리합니다.

5. 학습/검증 데이터 분리

- **분리 방법**: 본 프로젝트는 모델 학습보다는 교육 콘텐츠 제공이 목적이므로, 데이터를 난이도별(Step 1~5)로 분할하여 배치합니다.
- **비율**: 각 트랙별로 입문(Level 1-2), 중급(Level 3), 고급(Level 4-5) 비중을 4:3:3 수준으로 유지하도록 데이터를 구성합니다.

6. 전처리 결과 요약 및 평가

- **정합성**: 전문가 리뷰와 JSON 스키마 검증을 통해 실무와 유사한 고품질 교육 콘텐츠 확보
- **안정성**: Soft Delete 전략 및 암호화 적용으로 데이터 복구 가능성 및 보안성 향상