Fine-tuning용 데이터셋 수집 방안

1. 공개 데이터셋 수집 방안

1.1 유치원 단계 **(5-7**세**)**

1.1.1 언어 발달 데이터

- **CHILDES** 데이터베이스: 세계 최대 규모의 아동 언어 발달 데이터
- **Common Voice**: Mozilla의 다국어 음성 인식 데이터
- **LibriVox**: 무료 오디오북 데이터 (동화, 동요)
- **Wikipedia Kids**: 아동용 위키피디아 콘텐츠
- 국립국어원 아동 언어 코퍼스: 한국어 아동 언어 발달 자료

1.1.2 수학 기초 개념 데이터

- **Khan Academy Kids**: 기초 수학 학습 콘텐츠
- **Counting Collections**: 수 개념 학습 자료
- **PBS Kids Math Games**: 교육용 수학 게임 데이터
- **Math Learning Center**: 조작 가능한 수학 도구 데이터

1.1.3 과학 탐구 데이터

- **NASA Kids Club**: 우주과학 교육 자료
- **National Geographic Kids**: 자연과학 사진 및 영상
- **Smithsonian's History Explorer**: 박물관 교육 자료
- **OpenStax CNX**: 무료 과학 교육 콘텐츠

1.2 초등학교 단계 **(8-13**세**)**

1.2.1 국어 교육 데이터

- **Project Gutenberg**: 고전 문학 작품 텍스트
- 국립중앙도서관 디지털 컬렉션: 한국 아동 도서
- **Sejong Corpus**: 세종 말뭉치 (한국어 언어 자료)
- **CommonLit**: 독해력 향상을 위한 지문 모음
- **ReadWorks**: 등급별 읽기 자료

1.2.2 수학 교육 데이터

- **Khan Academy**: 초등 수학 전 과정
- **IXL Learning**: 수학 연습 문제 데이터
- **Math Playground**: 수학 게임 및 문제 해결 자료
- **NCTM Illuminations**: 미국 수학교사협의회 교육 자료
- **GeoGebra**: 수학 시각화 도구 및 활동

1.2.3 과학 교육 데이터

- **PhET Interactive Simulations**: 과학 시뮬레이션
- **NASA Education**: 우주과학 교육 프로그램
- **NOAA Education Resources**: 기상 및 해양과학 자료
- **Exploratorium**: 과학 박물관 교육 콘텐츠
- **SciShow Kids**: 과학 교육 영상 자료

1.2.4 사회 교육 데이터

- **National Geographic Education**: 지리 및 문화 교육 자료
- **Smithsonian Learning**: 역사 및 문화 교육 콘텐츠
- **iCivics**: 시민 교육 게임 및 자료
- 교육부 사회과 교육과정: 한국 사회과 표준 교육과정

1.3 중학교 단계 **(14-16**세**)**

1.3.1 국어 교육 데이터

- **Korean Literature Corpus**: 한국 문학 작품 데이터베이스
- **Naver** 지식백과: 문학 및 언어 관련 백과사전
- 한국고전번역원: 고전 문학 번역 자료
- 교보문고 **eBook**: 청소년 문학 작품 (저작권 해결 시)

1.3.2 수학 교육 데이터

- **Wolfram MathWorld**: 수학 개념 및 정리 데이터베이스
- **MIT OpenCourseWare**: 중등 수학 강의 자료
- **Khan Academy Algebra**: 대수학 교육 콘텐츠
- **GeoGebra Classroom**: 중학 수학 활동 자료

1.3.3 과학 교육 데이터

- **CK-12 Foundation**: 무료 과학 교과서 및 자료
- **MIT Blossoms**: 과학 교육 영상 강의
- **NOVA Education**: PBS의 과학 교육 프로그램
- 국립과천과학관: 과학 전시 및 교육 자료

1.3.4 사회 교육 데이터

- **Stanford History Education Group**: 역사 교육 자료
- **National Archives**: 역사 문서 및 자료
- **World Bank Open Data**: 세계 통계 및 경제 데이터
- **UN Data**: 국제기구 통계 자료

1.4 고등학교 단계 **(17-19**세**)**

1.4.1 심화 학습 데이터

- **MIT OpenCourseWare**: 대학 수준 강의 자료

- **Stanford Online**: 온라인 강의 콘텐츠
- **Coursera Public Courses**: 무료 온라인 강의
- **edX**: 세계 유명 대학의 온라인 강의

1.4.2 대학 입시 준비 데이터

- **College Board**: SAT 및 AP 시험 자료
- 한국교육과정평가원: 수능 기출문제 및 해설
- **ETS**: TOEFL 및 영어 능력 평가 자료

1.5 대학교 단계 **(20-23**세**)**

1.5.1 전공별 학술 데이터

- **arXiv**: 물리학, 수학, 컴퓨터과학 논문 저장소
- **PubMed**: 의학 및 생명과학 논문 데이터베이스
- **IEEE Xplore**: 공학 및 기술 논문 데이터베이스
- **JSTOR**: 인문학 및 사회과학 학술 자료
- **Google Scholar**: 학술 검색 엔진 데이터

1.5.2 교육 과정 데이터

- **Open Yale Courses**: 예일대학교 공개 강의
- **Harvard Extension School**: 하버드 확장 교육 프로그램
- **Berkeley Webcasts**: UC 버클리 강의 영상
- **KOCW**: 한국 대학의 공개 강의

2. 비공개 데이터셋 수집 방안

2.1 유치원 단계 **(5-7**세**)**

2.1.1 실제 교실 수업 데이터

- 유치원 협력 프로그램: 전국 유치원과 파트너십 구축
- 교사**-**아동 상호작용 녹음: IRB 승인 하 수업 대화 수집
- 놀이 활동 관찰: 자유놀이 시간의 학습 상황 기록
- 부모**-**자녀 학습 일지: 가정에서의 학습 과정 수집

2.1.2 발달 평가 데이터

- 인지 발달 평가: 전문 기관과 협력한 발달 검사 결과
- 언어 발달 추적: 개별 아동의 언어 사용 변화 기록
- 사회성 발달 관찰: 또래 상호작용 패턴 분석

2.2 초등학교 단계 **(8-13**세**)**

2.2.1 학교 교육과정 데이터

- 교육청 협력 프로그램: 지역 교육청과 데이터 수집 협약
- 수업 녹화 및 전사: 실제 수업의 교사-학생 대화 수집
- 학습지 및 시험 답안: 학생들의 실제 학습 결과물
- 개별 학습 진단: 학습 부진 및 우수 학생 사례 분석

2.2.2 학습자 반응 데이터

- 오답 패턴 분석: 학생들이 자주 틀리는 문제 유형 수집
- 학습 질문 데이터: 학생들이 수업 중 하는 질문 모음
- 과제 수행 과정: 숙제나 프로젝트 진행 과정 기록

2.3 중학교 단계 **(14-16**세**)**

2.3.1 학습 성취도 데이터

- 중학교 네트워크 구축: 전국 중학교와 연구 협력
- 정기 평가 결과: 중간고사, 기말고사 성적 및 문제 분석
- 수행평가 자료: 프로젝트, 발표, 실험 보고서 등
- 진로 탐색 활동: 학생들의 관심 분야 및 적성 검사 결과

2.3.2 심화 학습 데이터

- 영재교육원 자료: 수학, 과학 영재 교육 프로그램 데이터
- 경시대회 문제: 각종 학업 경시대회 문제 및 해답
- 자기주도학습 일지: 학생들의 개별 학습 계획 및 실행 기록

2.4 고등학교 단계 **(17-19**세**)**

2.4.1 입시 준비 데이터

- 고등학교 교육과정: 문·이과 선택 과목별 학습 자료
- 모의고사 결과: 전국 모의고사 문제 및 학생 답안 분석
- 대학 입시 상담: 진로 상담교사의 학생 지도 사례
- 논술 및 면접 자료: 대학별 전형 준비 과정 기록

2.4.2 심화 전공 탐색 데이터

- 과학고 교육과정: 과학고등학교의 심화 과학 교육 자료
- 외국어고 프로그램: 외국어 특성화 교육 과정
- 예술고 작품집: 예술 분야 특성화 교육 결과물

2.5 대학교 단계 **(20-23**세**)**

2.5.1 대학 교육과정 데이터

- 대학 강의 녹화: 주요 대학과 협력한 강의 수집
- 과제 및 시험 답안: 학생들의 실제 학습 결과물 분석
- 연구 프로젝트: 학부생 연구 과정 및 결과 수집
- 실습 및 인턴십: 현장 실무 경험 기록

2.5.2 전공별 전문 데이터

- 의과대학 임상 사례: 의학 교육용 임상 케이스 스터디
- 공과대학 설계 프로젝트: 엔지니어링 설계 과정 기록
- 인문대학 논문 작성: 인문학 연구 방법론 및 논문 사례
- 사회과학 데이터 분석: 사회과학 연구 프로젝트 수집

3. 데이터 수집 시 고려사항

3.1 윤리적 고려사항

- 개인정보 보호: 모든 개인 식별 정보 제거 및 익명화
- **IRB** 승인: 기관생명윤리위원회 심의 통과
- 동의서 취득: 학습자 및 보호자의 명시적 동의
- 데이터 보안: 암호화 및 접근 권한 관리

3.2 품질 관리

- 데이터 검증: 교육 전문가의 내용 검토
- 표준화 작업: 일관된 형식으로 데이터 변환
- 오류 제거: 잘못된 정보나 부적절한 내용 필터링
- 지속적 업데이트: 교육과정 변화에 따른 데이터 갱신

3.3 법적 고려사항

- 저작권 처리: 저작권 보호 자료의 적법한 사용
- 라이선스 준수: 공개 데이터의 사용 조건 확인
- 계약 체결: 비공개 데이터 제공 기관과의 협약
- 국제 규정: GDPR 등 국제적 데이터 보호 규정 준수

4. 데이터 처리 및 관리 체계

4.1 데이터 분류 체계

- 교육 단계별 분류: 유치원부터 대학까지 체계적 분류
- 과목별 분류: 국어, 수학, 과학, 사회, 예체능 등
- 난이도별 분류: 기초, 중급, 고급 수준별 구분
- 유형별 분류: 이론, 실습, 평가, 상호작용 등

4.2 데이터베이스 구축

- 통합 데이터베이스: 모든 수집 데이터의 중앙 집중식 관리
- 메타데이터 관리: 데이터 출처, 수집 일시, 품질 수준 기록
- 검색 시스템: 효율적인 데이터 검색 및 활용 도구
- 백업 시스템: 데이터 손실 방지를 위한 다중 백업

4.3 데이터 활용 전략

- 단계별 **Fine-tuning**: 교육 단계에 맞는 데이터 선별 사용
- 과목별 특화: 각 과목의 특성에 맞는 데이터 조합
- 개인화 학습: 학습자 특성에 따른 맞춤형 데이터 활용
- 지속적 개선: 사용 결과 피드백을 통한 데이터 품질 향상