DeepSeek R1 Fine-tuning 데이터셋 상세 명세서

제주한의약연구원 **AI Agent** 개발용

1. 개요

DeepSeek R1은 복잡한 추론과 다단계 사고를 요구하는 작업에 특화된 모델로,
제주한의약연구원에서는 다음 두 가지 전문 분석 Agent에 활용됩니다:

1. 약물 상호작용 분석 **Agent**: 한약-양약 상호작용, 복합 처방 최적화
2. 연구 데이터 분석 **Agent**: 실험 데이터 통계 분석, 메타분석, 연구 인사이트 도출

2. 공개 데이터셋

2.1 한의학 고전 문헌 데이터

A. 동의보감 데이터베이스

- 출처: 한국한의학연구원 한의학고전DB
- 내용:
    - 처방 구성 및 효능: 약 4,000개 처방
    - 약재별 성미귀경 정보: 1,400여 종 약재
    - 병증별 치료 원칙: 5,000여 개 병증
- 형식: JSON, XML
- 데이터량: 약 50,000개 레코드
- 활용: 전통 처방 분석 및 약물 조합 패턴 학습

B. 방약합편 데이터

- 출처: 한국한의학연구원
- 내용:
    - 상용 처방 572개
    - 처방 구성 원칙 및 가감법
    - 약재 배합 금기 사항
- 데이터량: 약 10,000개 텍스트 세그먼트
- 활용: 약물 배합 규칙 학습

C. 본초강목 데이터

- 출처: 중국 중의약 데이터베이스
- 내용:
    - 1,892종 약재 정보

    - 약재별 화학성분 (현대 연구 추가)
    - 약리작용 및 임상응용
- 데이터량: 약 30,000개 항목
- 활용: 약재 특성 및 작용 메커니즘 이해

2.2 현대 한의학 연구 데이터

A. PubMed 한의학 논문

- 출처: PubMed Central
- 검색 키워드: "Korean medicine", "Traditional Korean medicine", "Herbal medicine", "Acupuncture"
- 기간: 2010-2024
- 내용:
    - 임상시험 결과: 약 3,000편
    - 기초연구 논문: 약 5,000편
    - 체계적 문헌고찰: 약 500편
- 형식: 초록 및 전문(오픈액세스)
- 데이터량: 약 100,000개 문단
- 활용: 최신 연구 동향 및 과학적 근거 학습

B. 한국한의학연구원 공개 데이터

- 출처: KIOM 데이터포털
- 내용:
    - 한약재 표준화 데이터
    - HPLC 지문 데이터
    - 안전성 평가 결과
- 데이터량: 약 20,000개 실험 결과
- 활용: 품질관리 및 표준화 방법론 학습

C. 식약처 한약재 안전성 데이터

- 출처: 식품의약품안전처
- 내용:
    - 한약재 중금속 기준치
    - 잔류농약 허용기준
    - 미생물 한도 기준
    - 독성 시험 결과
- 데이터량: 약 5,000개 규정 및 시험 데이터
- 활용: 안전성 평가 기준 학습

2.3 약물 상호작용 데이터베이스

A. 한약**-**양약 상호작용 **DB**

- 출처: 대한한의사협회, 건강보험심사평가원
- 내용:
    - 문헌 보고된 상호작용 사례: 2,000건

    - 약동학적 상호작용 메커니즘
    - 임상적 중요도 분류
- 형식: 구조화된 데이터베이스
- 데이터량: 약 10,000개 상호작용 쌍
- 활용: 약물 상호작용 예측 모델 학습

B. DrugBank 데이터

- 출처: DrugBank 공개 데이터
- 내용:
    - 약물 화학구조 정보
    - 약물 대사 경로
    - CYP450 효소 상호작용
- 데이터량: 약 15,000개 약물 프로파일
- 활용: 약물 대사 메커니즘 이해

2.4 통계 분석 교육 데이터

A. 의학통계 교과서 데이터

- 출처: 공개된 의학통계 교재
- 내용:
    - 임상시험 설계 방법론
    - 통계 분석 기법 설명
    - 실제 분석 사례
- 데이터량: 약 50,000개 예제 및 설명
- 활용: 통계 분석 방법론 학습

B. R/Python 통계 분석 코드

- 출처: GitHub, Kaggle
- 내용:
    - 의학 데이터 분석 스크립트
    - 메타분석 코드 예제
    - 시각화 템플릿
- 데이터량: 약 10,000개 코드 스니펫
- 활용: 실제 분석 구현 방법 학습

2.5 제주 자생 약용식물 데이터

A. 제주 생물종다양성연구소 데이터

- 출처: 제주특별자치도 생물종다양성연구소
- 내용:
    - 제주 자생 약용식물 500종
    - 서식지 및 분포 정보
    - 형태학적 특징
- 데이터량: 약 10,000개 식물 정보
- 활용: 제주 특화 약재 이해

B. 국립수목원 약용식물 **DB**

- 출처: 국립수목원
- 내용:
    - 한국 자생 약용식물 정보
    - 전통 이용 방법
    - 현대 연구 결과
- 데이터량: 약 8,000개 식물 프로파일
- 활용: 약용식물 활용 방안 학습

3. 비공개 데이터셋

3.1 제주한의약연구원 내부 연구 데이터

A. 연구 프로젝트 데이터

- 내용:
    - 진행중인 연구과제 실험 데이터: 50개 과제
    - 실험 프로토콜 및 SOP: 200개 문서
    - 연구노트 및 실험 기록: 10,000페이지
    - 내부 세미나 발표 자료: 500개 프레젠테이션
- 보안 등급: 기밀
- 데이터량: 약 100,000개 실험 결과
- 처리 방법:
    - 연구자 실명 제거
    - 특허 출원 예정 내용 마스킹
    - 공동연구 기관 정보 익명화

B. 임상시험 데이터

- 내용:
    - IRB 승인 임상시험 데이터: 20개 시험
    - 환자 증례 기록: 2,000명 (익명화)
    - 이상반응 보고서: 500건
    - 효능 평가 결과: 5,000개 측정값
- 보안 등급: 극비
- 데이터량: 약 50,000개 임상 데이터
- 처리 방법:
    - 완전 익명화 (k-anonymity 적용)
    - 민감 정보 제거
    - 통계 요약 데이터만 활용

C. 분석 장비 원시 데이터

- 내용:
    - HPLC 크로마토그램: 50,000개
    - LC-MS/MS 스펙트럼: 30,000개
    - NMR 데이터: 10,000개

    - 현미경 이미지: 20,000장
- 보안 등급: 대외비
- 데이터량: 약 110,000개 분석 결과
- 처리 방법:
    - 메타데이터만 추출
    - 피크 패턴 정보 활용
    - 이미지는 특징값만 추출

3.2 협력 기관 공유 데이터

A. 제주대학교병원 한방진료 데이터

- 내용:
    - 한방 처방 기록: 50,000건 (익명화)
    - 진료 경과 기록: 100,000건
    - 검사 결과 연계 데이터: 30,000건
- 보안 등급: 기밀
- 데이터량: 약 180,000개 진료 기록
- 활용 조건:
    - IRB 승인 필요
    - 병원 데이터는 병원 서버에서만 처리
    - 학습된 모델만 반출 가능

B. 제주 한의약 기업 생산 데이터

- 내용:
    - 제품 품질 검사 데이터: 10,000 배치
    - 제조 공정 파라미터: 50,000개 측정값
    - 안정성 시험 결과: 5,000개 시험
- 보안 등급: 영업비밀
- 데이터량: 약 65,000개 품질 데이터
- 처리 방법:
    - 기업명 및 제품명 코드화
    - 핵심 제조 기술 제외
    - 일반화된 패턴만 학습

3.3 전문가 지식 데이터

A. 한의학 전문가 상담 기록

- 내용:
    - 연구원 자문 회의록: 1,000건
    - 전문가 Q&A 세션: 5,000개 질답
    - 처방 검토 의견서: 2,000건
- 보안 등급: 내부용
- 데이터량: 약 8,000개 전문가 의견
- 처리 방법:
    - 전문가 신원 보호

    - 핵심 노하우만 추출
    - 일반화 가능한 지식으로 변환

B. 특허 출원 준비 자료

- 내용:
    - 특허 명세서 초안: 100건
    - 선행기술 조사 보고서: 200건
    - 특허 전략 문서: 50건
- 보안 등급: 극비
- 데이터량: 약 350개 문서
- 처리 방법:
    - 핵심 청구항 내용 제외
    - 일반적 특허 작성 패턴만 학습
    - 공개 특허와 비교하여 차별점만 활용

3.4 실패 실험 및 부정적 결과 데이터

A. 실패한 실험 기록

- 내용:
    - 목표 미달성 실험: 3,000건
    - 실패 원인 분석: 1,000건
    - 개선 방안 보고서: 500건
- 보안 등급: 내부용
- 데이터량: 약 4,500개 실패 사례
- 가치:
    - 잘못된 접근 방법 학습
    - 문제 해결 패턴 이해
    - 연구 방향 최적화

B. 부작용 및 안전성 이슈 데이터

- 내용:
    - 예상치 못한 부작용 사례: 200건
    - 독성 시험 양성 결과: 100건
    - 품질 이탈 사례: 300건
- 보안 등급: 기밀
- 데이터량: 약 600개 안전성 이슈
- 처리 방법:
    - 제품 식별 정보 완전 제거
    - 패턴과 메커니즘만 추출
    - 예방적 안전성 평가에 활용

4. 데이터 전처리 및 품질 관리

4.1 데이터 정제 프로세스

A. 공개 데이터 처리

1. 형식 표준화
    - 모든 텍스트를 UTF-8로 인코딩
    - 한의학 용어 표준화 (표준 용어집 적용)
    - 약재명 통일 (이명 처리)
2. 품질 검증
    - 중복 데이터 제거
    - 불완전한 레코드 보완 또는 제외
    - 출처 신뢰도 평가
3. 구조화
    - JSON-LD 형식으로 변환
    - 메타데이터 추가
    - 관계형 데이터 연결

B. 비공개 데이터 처리

1. 익명화 및 가명화
    - 개인정보 완전 제거
    - k-익명성 (k≥5) 보장
    - 차분 프라이버시 적용
2. 보안 처리
    - 데이터 암호화 (AES-256)
    - 접근 권한 관리
    - 감사 로그 유지
3. 민감도 분류
    - 5단계 보안 등급 적용
    - 등급별 처리 규칙 적용
    - 외부 반출 가능 데이터 선별

4.2 데이터 증강 기법

A. 텍스트 데이터 증강

1. 패러프레이징
    - 동일 의미의 다양한 표현 생성
    - 한의학 용어 동의어 치환
    - 문장 구조 변형
2. 역번역 **(Back-translation)**
    - 한국어 → 영어 → 한국어

    - 한국어 → 중국어 → 한국어
    - 의미 보존 검증
3. 컨텍스트 확장
    - 관련 배경 지식 추가
    - 인과관계 명시화
    - 추론 과정 상세화

B. 수치 데이터 증강

1. 노이즈 추가
    - 가우시안 노이즈 (σ=0.05)
    - 측정 오차 범위 내 변동
    - 이상치 생성 및 처리
2. 보간법 **(Interpolation)**
    - 실험 조건 간 중간값 생성
    - 용량-반응 곡선 보간
    - 시계열 데이터 확장
3. 시뮬레이션 데이터
    - PBPK 모델링 결과
    - 몬테카를로 시뮬레이션
    - 약동학 파라미터 예측

4.3 데이터셋 구성 비율

Fine-tuning 데이터셋 최종 구성

- 공개 데이터: 60% (약 300,000개 샘플)
    - 한의학 고전 문헌: 15%
    - 현대 연구 논문: 20%
    - 약물 상호작용 DB: 15%
    - 통계 분석 교육 자료: 10%
- 비공개 데이터: 40% (약 200,000개 샘플)
    - 내부 연구 데이터: 15%
    - 임상/진료 데이터: 10%
    - 전문가 지식: 10%
    - 실패 사례 및 안전성: 5%

검증 데이터셋

- 총량: 전체의 20% (100,000개 샘플)
- 구성: 공개/비공개 동일 비율
- 용도: 과적합 방지, 성능 평가

테스트 데이터셋

- 총량: 전체의 10% (50,000개 샘플)
- 구성: 실제 업무 시나리오 중심
- 용도: 최종 성능 검증

5. 데이터 거버넌스

5.1 데이터 수집 및 활용 원칙

1. 적법성
    - 모든 데이터 수집은 관련 법규 준수
    - 저작권 및 라이선스 확인
    - 개인정보보호법 준수
2. 윤리성
    - IRB 승인 획득 (임상 데이터)
    - 정보 주체 동의 확보
    - 연구 윤리 준수
3. 투명성
    - 데이터 출처 명시
    - 처리 과정 문서화
    - 활용 목적 공개

5.2 데이터 보안 체계

1. 물리적 보안
    - 전용 서버실 운영
    - 출입 통제 시스템
    - CCTV 모니터링
2. 기술적 보안
    - 데이터 암호화
    - 접근 권한 관리
    - 네트워크 격리
3. 관리적 보안
    - 보안 교육 정기 실시
    - 데이터 취급자 서약서
    - 정기 보안 감사

5.3 데이터 생명주기 관리

1. 수집 단계
    - 출처 검증
    - 품질 확인
    - 메타데이터 생성
2. 처리 단계
    - 버전 관리
    - 변경 이력 추적
    - 백업 체계 운영
3. 활용 단계
    - 접근 로그 관리
    - 사용 목적 제한
    - 성과 측정
4. 폐기 단계
    - 보존 기간 준수
    - 안전한 폐기
    - 폐기 증명서 발급

6. 예상 성과 및 품질 지표

6.1 데이터셋 품질 지표

- 완전성: **95%** 이상
- 정확성: **98%** 이상
- 일관성: **99%** 이상
- 시의성: 최신 데이터 비율 **80%** 이상

6.2 모델 성능 목표

- 약물 상호작용 예측 정확도: **92%** 이상
- 통계 분석 적절성: **95%** 이상
- 추론 과정 설명가능성: **90%** 이상
- 응답 시간: 평균 **2**초 이내

6.3 보안 및 컴플라이언스

- 데이터 유출 사고: **0**건
- 개인정보 노출: **0**건
- 감사 지적사항: 연간 **2**건 이하
- 보안 인증: **ISO 27001** 획득

이 데이터셋 구성은 제주한의약연구원의 AI Agent가 높은 전문성과 신뢰성을 갖추면서도, 데이터 보안과 윤리적 기준을 철저히 준수할 수 있도록 설계되었습니다.