제주국제자유도시개발센터(JDC) AI Agent 개발 계획

1. JDC AI Agent 개발용 데이터셋 분류표

📊 공개 데이터셋 (Public Datasets)

카테고리	데이터명	출처	용도	데이터 규모
법령/규정	JDC 설립 및 운영에 관한 특별법	국가법령정보센터	기본 법적 근거 학습	50페이지
	제주국제자유도시 특별법	국가법령정보센터	법적 프레임워크 이해	200페이지
	도시개발법령집	국토교통부	개발 관련 법규 학습	500페이지
개발사업	JDC 공시 사업계획서	JDC 홈페이지	사업 이해 및 안내	1,000건
	준공된 개발사업 현황	JDC 연차보고서	사업 성과 학습	300건
	도시계획 변경 공고문	제주도청 홈페이지	계획 변경 절차	200건
분양/임대	분양공고문 (과거 5년)	JDC 홈페이지	분양 절차 및 조건	150건
	임대사업 안내자료	JDC 홈페이지	임대 관련 업무	80건
	입주기업 현황	공개 보도자료	입주 절차 안내	500건
관광/레저	관광단지 조성계획	문화체육관광부	관광 개발 이해	100건
	골프장 운영현황	JDC 공시자료	시설 운영 안내	50건

	컨벤션센터 운영실적	ICC JEJU 홈페이지	시설 이용 안내	200건
교통/인프라	대중교통 연계계획	제주도청 교통정책과	교통 안내	100건
	상하수도 연계현황	제주상하수도본부	인프라 정보	80건
환경/안전	환경영향평가서	환경부 환경영향평가정보지원시스템	환경 규제 이해	50건
	안전관리계획서	고용노동부 안전보건공단	안전 절차 학습	30건
일반행정	조직도 및 업무분장	JDC 홈페이지	조직 이해	20페이지
	공지사항 (5년간)	JDC 홈페이지	일반 안내 업무	2,000건

공개 데이터셋 총계**:** 약 **5,490**건 **/ 970**페이지

🔒 비공개 데이터셋 **(Private/Internal Datasets)**

카테고리	데이터명	보유부서	용도	데이터 규모
민원처리	민원접수/처리 이력 (3년간)	고객지원팀	민원 응답 패턴 학습	15,000건
	전화상담 녹취록	고객지원팀	대화형 AI 학습	5,000시간
	온라인 문의답변 DB	IT팀	자동응답 학습	8,000건
내부업무	사업승인 프로세스 문서	사업기획팀	내부 절차 자동화	500건
	계약관리 시스템 데이터	계약관리팀	계약 관련 업무	2,000건
	예산집행 내역	재무팀	재무 관련 문의	1,000건
고객정보	분양고객 문의유형 분석	분양팀	맞춤형 상담	3,000건
	임대고객 요구사항 DB	임대팀	서비스 개선	1,500건
	컨벤션 이용고객 피드백	컨벤션운영팀	시설 안내 최적화	2,000건
운영현황	시설별 이용현황 통계	시설관리팀	실시간 정보 제공	365일×5년

	교통량 및 주차현황	교통관리팀	교통 안내 최적화	365일×3년
	보안/출입 관리 기록	보안팀	보안 절차 안내	365일×3년
기술정보	설계도면 및 시방서	건설기술팀	기술 문의 대응	10,000건
	시공 및 감리 보고서	건설관리팀	전문 기술 상담	5,000건
	유지보수 이력	시설관리팀	시설 관리 안내	8,000건
법무/규정	내부 업무규정집	법무팀	내부 절차 안내	200건
	계약서 표준양식	법무팀	계약 업무 지원	100건
	분쟁처리 사례집	법무팀	분쟁 예방 상담	300건

비공개 데이터셋 총계**:** 약 **61,965**건 **+** 연속 데이터 **3-5**년

2. DeepSeek R1 파인튜닝 일일 작업 계획

🎯 선정 모델**: DeepSeek-R1-32B (JDC** 메인 **AI**용**)**

📅 **1**주차**:** 환경 구축 및 데이터 준비 **(Day 1-7)**

Day 1 (2025.01.13) - 환경 설정

- 오전 **(09:00-12:00)**
  - 서버 환경 구축 (A100 40GB × 4 GPU 클러스터)
  - CUDA 12.1, PyTorch 2.1, Transformers 라이브러리 설치
  - DeepSeek-R1-32B 베이스 모델 다운로드 및 테스트
- 오후 **(13:00-18:00)**
  - 파인튜닝 환경 설정 (LoRA, QLoRA 설정)
  - Git repository 설정 및 코드 버전 관리 시스템 구축
  - 모델 로딩 및 기본 추론 테스트

Day 2 (2025.01.14) - 공개 데이터 수집

- 오전 **(09:00-12:00)**
  - JDC 홈페이지 크롤링 (공지사항, 사업계획서)
  - 제주도청 관련 공시자료 수집
  - 법령 데이터 다운로드 및 정리
- 오후 **(13:00-18:00)**
  - 수집된 데이터 품질 검증
  - 중복 데이터 제거 및 형식 통일
  - 데이터 분류 및 라벨링 작업 시작

Day 3 (2025.01.15) - 비공개 데이터 접근

- 오전 **(09:00-12:00)**
  - JDC 내부 시스템 접근 권한 확보
  - 민원처리 시스템 데이터 추출
  - 개인정보 마스킹 및 익명화 처리
- 오후 **(13:00-18:00)**
  - 상담 녹취록 텍스트 변환 (STT)
  - 내부 업무 문서 디지털화
  - 데이터 보안 정책 수립 및 적용

Day 4 (2025.01.16) - 데이터 전처리 **(1**차**)**

- 오전 **(09:00-12:00)**
  - 텍스트 정규화 (한글 맞춤법, 띄어쓰기)
  - 불필요한 메타데이터 제거
  - 문서 구조 파싱 (제목, 본문, 첨부 구분)
- 오후 **(13:00-18:00)**
  - 데이터 품질 점검 스크립트 실행
  - 이상치 데이터 식별 및 처리
  - 데이터 통계 분석 (길이, 빈도, 패턴)

Day 5 (2025.01.17) - 데이터 전처리 **(2**차**)**

- 오전 **(09:00-12:00)**
  - Question-Answer 페어 생성
  - 대화형 데이터 포맷 변환
  - 컨텍스트-응답 매칭 작업
- 오후 **(13:00-18:00)**
  - 훈련/검증/테스트 데이터셋 분할 (8:1:1)
  - 데이터 밸런싱 (카테고리별 균등 분배)
  - 최종 데이터셋 검증 및 저장

Day 6 (2025.01.18) - 파인튜닝 설정

- 오전 **(09:00-12:00)**
  - 훈련 하이퍼파라미터 설정
    - Learning rate: 5e-5
    - Batch size: 4 (gradient accumulation: 8)
    - Max sequence length: 4096
    - LoRA rank: 64, alpha: 128
- 오후 **(13:00-18:00)**
  - 평가 메트릭 설정 (BLEU, ROUGE, Perplexity)
  - 체크포인트 저장 전략 수립
  - 로그 및 모니터링 시스템 구축

Day 7 (2025.01.19) - 파일럿 테스트

- 오전 **(09:00-12:00)**

    - 소규모 데이터셋으로 파일럿 훈련 (1,000샘플)
    - 훈련 파이프라인 검증
    - 메모리 사용량 및 속도 체크
- 오후 **(13:00-18:00)**
    - 파일럿 결과 분석
    - 하이퍼파라미터 미세 조정
    - 본격 훈련 준비 완료

📅 **2**주차**:** 기본 파인튜닝 **(Day 8-14)**

Day 8 (2025.01.20) - 1차 파인튜닝 시작

- 오전 **(09:00-12:00)**
    - 전체 데이터셋 1차 훈련 시작 (Epoch 1)
    - 실시간 모니터링 및 로그 확인
    - GPU 사용률 및 온도 체크
- 오후 **(13:00-18:00)**
    - 중간 체크포인트 저장 및 평가
    - Loss curve 분석
    - 메모리 최적화 조정

Day 9-11 (2025.01.21-23) - 지속적 훈련

- 매일 동일 일정
    - 오전: 훈련 모니터링, 이상 상황 대응
    - 오후: 중간 평가, 샘플 응답 품질 체크
    - 저녁: 다음 날 훈련 계획 수립

Day 12 (2025.01.24) - 1차 평가

- 오전 **(09:00-12:00)**
    - 1차 훈련 완료 및 모델 저장
    - 검증 데이터셋으로 성능 평가
    - 정량적 지표 측정 (BLEU, ROUGE)
- 오후 **(13:00-18:00)**
    - 정성적 평가 (실제 민원 시뮬레이션)
    - 응답 품질 및 정확성 검토
    - 개선 포인트 도출

Day 13-14 (2025.01.25-26) - 1차 개선

- **Day 13**: 하이퍼파라미터 재조정, 문제 데이터 재처리
- **Day 14**: 2차 훈련 준비, 추가 데이터 수집

📅 **3**주차**:** 전문화 파인튜닝 **(Day 15-21)**

Day 15-17 (2025.01.27-29) - 도메인 특화 훈련

- **Day 15**: JDC 전문 용어 및 업무 프로세스 집중 훈련

- **Day 16**: 민원 응답 패턴 특화 훈련
- **Day 17**: 다국어 지원 (영어, 중국어) 추가 훈련

Day 18-19 (2025.01.30-31) - 대화형 **AI** 최적화

- **Day 18**: 멀티턴 대화 성능 개선
- **Day 19**: 컨텍스트 유지 능력 강화

Day 20-21 (2025.02.01-02) - 2차 평가 및 조정

- **Day 20**: 종합 성능 평가
- **Day 21**: 최종 하이퍼파라미터 최적화

📅 **4**주차**:** 통합 테스트 및 배포 준비 **(Day 22-28)**

Day 22-24 (2025.02.03-05) - 시스템 통합

- **Day 22**: JDC 기존 시스템과 API 연동 테스트
- **Day 23**: 실시간 응답 속도 최적화
- **Day 24**: 보안 및 개인정보 보호 테스트

Day 25-26 (2025.02.06-07) - 사용자 테스트

- **Day 25**: JDC 직원 대상 베타 테스트
- **Day 26**: 피드백 수집 및 반영

Day 27-28 (2025.02.08-09) - 최종 배포

- **Day 27**: 최종 모델 검증 및 배포 패키지 생성
- **Day 28**: 운영 환경 배포 및 모니터링 시스템 가동

3. 기대 성과 및 평가 지표

📊 정량적 지표

- 응답 정확도: 95% 이상
- 응답 속도: 평균 2초 이내
- 고객 만족도: 90% 이상
- 업무 효율성: 70% 향상

🎯 정성적 지표

- 자연스러운 대화: 인간과 유사한 응답 품질
- 전문성: JDC 업무 영역 전문 지식 보유
- 일관성: 24시간 동일한 서비스 품질 유지
- 다국어 지원: 한국어, 영어, 중국어 지원

4. 리스크 관리 및 대응 방안

⚠️ 주요 리스크

- 데이터 품질 이슈: 지속적 모니터링 및 개선
- 모델 편향성: 다양한 평가 데이터셋으로 검증
- 시스템 안정성: 이중화 구성 및 백업 시스템
- 개인정보 보호: 엄격한 데이터 보안 정책 적용

🛡️ 대응 방안

- 일일 모니터링: 성능 지표 실시간 추적
- 주간 평가: 정기적 성능 평가 및 개선
- 월간 업데이트: 새로운 데이터로 지속적 학습
- 분기별 검토: 전체 시스템 점검 및 업그레이드