[DEEP DIVE] 1차 프로젝트 기획서

프로젝트 기간	2025.06.16(월) ~ 2025.07.04(금)		
일자	2025.06.16(월)		
과정명(회차)	생성 AI 응용 서비스 개발자 양성 과정 (3회차)		
참여인원	5명	**팀장**	최정훈
구성원	최정훈, 박지연, 이서준, 이재진, 안효서		
프로젝트명	AI OCR 기반 임대차 계약서 자동분석 및 전세사기 위험탐지 시스템 개발		
주제	임대차 계약서와 등기부등본을 **OCR**로 분석하여, 주소, 보증금, 특약 사항 등의 핵심 정보를 자동으로 추출하고, LLM 기반 자연어 처리 기법을 통해 **불공정 조항 및 전세사기 위험 요소를 판별**하며, 등기부등본의 표제부·갑구·을구 정보를 시각화함으로써 사용자에게 종합적인 **전세사기 위험도를 제공**하는 것을 목표로 함.		

주제 선정 이유	최근 주거 관련 사기, 특히 **전세사기 문제가 사회적으로 큰 이슈로 대두**되고 있으며, 이에 따라 임대차 계약서와 등기부등본 등 주요 문서에 대한 정확한 분석과 검토 수요가 급증하고 있음. 그러나 **일반 사용자**들은 이러한 문서의 법적·행정적 구조와 내용을 정확히 이해하기 어려우며, 위험 요소를 사전에 파악하는 것도 쉽지 않은 상황임. 특히, **임대차 계약서** 내 불공정 조항이나 등기부등본 상의 근저당권 설정 등 권리관계는 **확인에 시간이 오래 걸리고**, 법률적 해석이 요구되는 **어려운 영역**임. 본 프로젝트는 이러한 문제를 해결하기 위해, **오픈소스 OCR 기술을 활용**하여 임대차 계약서 및 등기부등본의 주요 항목(주소, 보증금, 특약사항, 소유자 정보 등)을 **자동 추출**하고, **생성형 AI(LLM)**를 통해 문서 내 약관 분석, **위험 요인 판단 및 전세사기 가능성 탐지** 기능을 구현하고자 함. 또한 추출된 정보를 시각화하여 사용자에게 직관적으로 제공함으로써, **임차인이 사전에 계약의 위험성을 인지하고, 안전한 거래를 유도할 수 있도록 지원**하고자 함. 이를 통해 전세사기에 대한 사회적 불안 요소를 완화하고, **불법 거래의 사전 예방 및 건전한 부동산 거래 문화 형성에 기여**하고자 함.

프로젝트 목표	**최종목표** 임대차 계약서 및 등기부등본 문서에 OCR과 후처리 기술을 적용해 전세사기 위험 요소를 자동으로 추출·분석하고, 이를 시각적으로 제공하는 AI 기반 서비스 프로토타입(MVP)을 개발하고자 함. **세부목표** **1. 도메인 선정** ● 주거/부동산 도메인의 임대차 계약서와 등기부등본 문서를 대상으로 설정함. ● 전세사기, 깡통전세 등 사회적 이슈가 지속됨에 따라 해당 문서에 대한 사전 검토 수요가 증가하고 있으며, 정형화된 문서 구조로 인해 OCR 및 AI 분석 기술 적용에 적합하다고 판단함. **2. 데이터 확보** ● 임대차 계약서 및 등기부등본 형식의 문서 30장 이상을 수집하고, 다양한 해상도와 서식, 서체를 포함하여 OCR 실험에 활용할 예정임. ● 실제 계약서와 등기부등본은 개인정보 보호 등의 이슈로 수집이 제한적이므로, 공개 사례를 참조하거나 실물 사례를 기반으로 유사 형식의 문서를 자체 제작하여 데이터셋을 구성할 예정임. ● 또한, 전세금과 시세 비교를 위한 국토교통부 실거래가 데이터를 연동하여 분석 기반을 구축할 예정임. **3. OCR 기술 적용** ● EasyOCR을 기반으로 문서 내 텍스트를 인식하고, 파인튜닝을 통해 문서 인식 정확도를 개선할 예정임. ● 특히 수기로 작성된 항목의 인식률 향상을 목표로 하여, 필기체 인식 정확도 80% 이상을 달성하는 것을 목표로 함. ● EasyOCR 결과를 Gemini 기반 후처리를 통해 오타를 교정하여 최종 정확도를 향상시킬 예정임. **4. 정보 추출** ● 임대차 계약서에서 주소, 전세금액(보증금), 월세금액, 관리비, 특약사항 등의 핵심 정보를 추출하여 JSON 또는 표 형태로 구조화할 예정임. ● 등기부등본은 표제부, 갑구, 을구 구간을 구분하여 주소, 임대인, 은행명, 채권최고액, 설정일 등을 추출하고 구조화할 예정임. ● 필기체 인식이 어려운 경우 사용자 입력을 통한 보완 기능을 적용하며, 전체 정보 추출 정확도는 95% 이상을 달성하는 것을 목표로 함. **5. LLM 기반 분석 기능** ● 추출된 계약서 및 등기부 정보를 바탕으로 생성형 AI(대형언어모델, LLM)를 활용하여 불공정 조항 여부 및 전세사기 위험 요소를 분석할 예정임.

	• 또한, 실거래가와 보증금의 비교 분석을 통해 전세가율 기반 위험도 분류 기능을 구현할 예정임. **6. UI 구현** • Streamlit을 활용하여 사용자가 문서를 업로드하고, 위험 분석 결과를 직관적으로 확인할 수 있는 웹 기반 프로토타입을 구현할 예정임. • 분석 결과는 아래와 같은 메시지 형식으로 제공하고자 함: **A.** 안전한 계약입니다 **B.** 주의: 이 주소는 사기 사례가 있는 지역입니다 **C.** 위험: 보증금 한도를 초과한 고위험 거래입니다 **7. 성능 평가** • OCR 성능은 문자 인식 정확도를 기준으로 측정할 예정임. • 정보 추출 성능은 Precision 및 Recall 지표를 활용하여 평가할 예정임. • 또한, 사용자 테스트를 통해 정성적 피드백을 수집하고 개선 방향을 도출할 예정임. **8. 확장성 기획** • Gemini 를 결합하여 계약서 전체 내용을 해석하고 요약하는 기능을 추가할 예정임. • 또한, 위험성이 높은 문장을 자동으로 하이라이팅하는 기능을 적용하여 사용자의 이해를 돕는 고도화 기능을 구현할 예정임.
기대효과	**기술적 기대 효과 (팀 역량 향상)** • OCR 오픈소스를 실습함으로써 다양한 문서 유형에 대한 인식 기술을 습득할 수 있음. • 이미지 전처리 및 데이터 정제 기법을 통해 문서 인식 정확도 향상 방안을 학습할 수 있음. • 정규표현식 및 템플릿 기반 정보 추출 자동화 경험을 통해 실무 적용 역량을 강화할 수 있음. • Gemini API 기반 LLM을 활용하여 약관 분석 및 자연어 처리 기술을 익힐 수 있음. • 국토교통부 실거래가 API 등 공공 데이터 연계 경험을 통해 외부 시스템 통합 역량을 향상시킬 수 있음. • OCR–LLM 파이프라인을 Streamlit으로 구현함으로써 웹 기반 AI 서비스 설계 경험을 쌓을 수 있음. • GitHub 및 Notion을 활용한 팀 기반 협업과 프로젝트 문서화 역량이 강화될 것으로 기대됨.**업무적 기대 효과 (서비스 가치)** • • 임대차 계약서 및 등기부등본 해석 자동화를 통해 비전문가도 직관적으로 계약 내용을 이해할 수 있을 것으로 기대됨. • 전세사기 위험 요소를 사전에 분석하고 시각화함으로써 사용자 불안을 줄이고 피해를 예방할 수 있음.

	● 불공정 조항 및 근저당권 탐지를 통해 정보 비대칭 문제를 해소할 수 있음. ● 계약 정보와 실거래가 데이터를 통합 분석함으로써 합리적인 계약 판단을 지원할 수 있음. ● 결과적으로 임차인의 재산권 보호 및 건전한 부동산 거래 문화 조성에 기여할 것으로 기대됨.
추진 일정	**1. 기획 (25.06.16 ~25.06.20)** **1.1** 도메인 설정 - 부동산 임대차 계약 및 등기부 등본 기반 전세사기 예방 도메인 확정 - 주요 문제 정의(보증금 과대, 불공정 특약, 명의 불일치 등) **1.2** 목표 및 지표 설정 - OCR 인식률(CER), 사용자 직관성(정성) 등 평가 지표 수립 - 성능 비교 실험을 통한 OCR 최종 선정 기준 마련 - Tesseract / EasyOCR 등 후보군 비교 계획 **1.3** 역할 분담 및 협업 체계 - OCR, 데이터 정제, LLM 분석, UI 개발 등 각 역할 분배 - GitHub/Notion 세팅, 회의 주기 및 정리 템플릿 구성 - OCR 벤치마크 실험 책임자 지정 및 테스트 조건 사전 설계 **2. 개발 (25.06.20 ~ 25.06.30)** **2.1** 전체 구조 및 파이프라인 설계 - 전체 서비스 흐름도 완성 - 모듈별 역할 정의 및 API 연동 흐름 확정 - 국토교통부 실거래가 API 실테스트 **2.2** 데이터 수집 및 전처리 - 임대차 계약서, 등기부등본 30장 수집 (직접 제작 포함) - 이미지 증강, 해상도 테스트, OCR 실험용 GT 데이터 구축 **2.3** OCR 성능 실험 및 튜닝 - Tesseract / EasyOCR 비교 실험 - 계약서 내 필기체 인식 여부, 도장/서명 영역 무시 처리 등 검토 - 전용 전처리 파이프라인 적용(OpenCV 등) **2.4** 정보 추출 및 시세 비교 모델 구현

	- 주소, 보증금, 특약, 계좌번호 등 필드별 정규표현식 기반 추출기 작성 - 실거래가 단지 평균(㎡당 시세) 계산 로직 개발 - 보증금 대비 시세 비율에 따른 위험도 분류 함수 작성 **2.5** LLM 분석 및 UI 개발 병행 - 불공정 특약 분석용 Prompt 설계 및 GPT API 연동 - Streamlit 기반 UI 제작(문서 업로드 → 결과 시각화 흐름 구성) - 예외 메시지 / 리스크 등급별 알림 메시지 작성 **3. 결과 정리 및 제출 (25.07.01 ~ 25.07.03)** **3.1** 결과 보고서 정리 - 프로젝트 구조도, OCR 정확도, 사용자 시나리오 등 포함 **3.2** 시연 영상 제작 - 실제 문서 업로드 → 추출 → 분석 → 판단 결과 흐름을 녹화 - 각 기능 설명 자막 포함 편집 **4. 최종 발표 (25.07.04)** **4.1** 최종 발표 - 문제 정의, 기술 구조 설명, 사용자 흐름 시연, 결과 및 기대효과 순 발표
역할분담	**최정훈 (팀장)** - 전체 일정관리 - 시스템 아키텍처 설계 - 개발 프로세스 총괄 - 문서 이미지 전처리 적용 - 인식률 비교 분석 및 후처리 최적화 **이재진 (실거래가 분석)** - 국토교통부 실거래가 API 연동 - 법정동코드 API 연동 - 동별 평균 거래금액, 단위 면적당 평균 거래금액 추출 **박지연 (LLM 기능 담당 및 문서 작성)** - 불공정 조항 탐지 - 전사세기 위험 분석을 위한 생성형 AI 모델 연동 - 문서 이미지 전처리 적용 - 인식률 비교 분석 및 최적화 **이서준 (UI/프론트 엔드 담당)** - Streamlit 기반 인터페이스 설계 및 구현

	- 문서 이미지 전처리 적용 - 인식률 비교 분석 및 최적화 **안효서 (AI OCR 기능 분석 및 인식률 개선 담당)** - OCR 정확도 향상을 위한 이미지 자동 보정 기능 구현 - 다양한 OCR 엔진 성능 비교 및 후처리(LLM 연동)를 통한 최종 텍스트 결과 정제 - 문서 이미지 전처리 적용 - 인식률 비교 분석 및 최적화