

유학생을 위한 중고거래 다국어 플랫폼 비즈니스 전략

당근마켓 크롤링 데이터 기반 가격 예측과 LLM 다국어 번역 지원 시스템

부산외국어대학교 | 6조 · 겐로그 | 정윤기 · 유준환 · 홍유진 · 김해나

01

주제 설정 배경 및 문제 정의

부산외국어대학교는 유학생 비중이 높은 대학으로, 유학생들은 당근마켓 등 중고거래 플랫폼을 자주 이용하지만 서비스가 한국어 전용으로 제공되어 이용에 어려움을 겪고, 게시글의 가격이 적정하지 판단하기 어렵다는 문제를 겪고 있다.

본 프로젝트는 유학생이 중고거래 시 겪는 언어 장벽과 가격 적정성 판단의 어려움을 해결하고, 중고거래 매물 데이터 기반 가격 예측과 다국어 통역을 함께 제공하는 것을 목표로 한다.

가설

중고거래 매물 데이터를 학습하면 적정 가격대를 예측할 수 있고, LLM 기반 번역 · 설명 생성을 결합하면 언어 장벽 문제도 함께 해결할 수 있다는 가설을 세웠다.

프로젝트 목표

실시간 번역
게시글과 채팅 메시지의 언어 장벽을 즉시 해소한다.

합리적 가격 예측
수집한 매물 데이터를 학습해 적정 가격대를 예측한다.

거래 지원
LLM 기반 번역 · 설명 생성으로 협상과 거래 과정을 지원한다.

성공 기준

- 중고거래 매물 100건 이상 크롤링 성공
- 가격 예측 모델 구축
- 번역 기능 에러율 0% (수집 매물 전건 정상 출력)
- 팀 자체 검증 20건 중 18건 이상 의미 정확 번역 (90% 이상)

02

데이터 수집



수집 방법

웹 스크래핑 (크롤링): Selenium(undetected_chromedriver)과 BeautifulSoup으로 당근마켓 부산광역시 100개 동네를 순회하며 게시글 · 가격 · 위치/시간 정보를 직접 수집했다 (동네당 최대 300건).

데이터 요약 (실제)

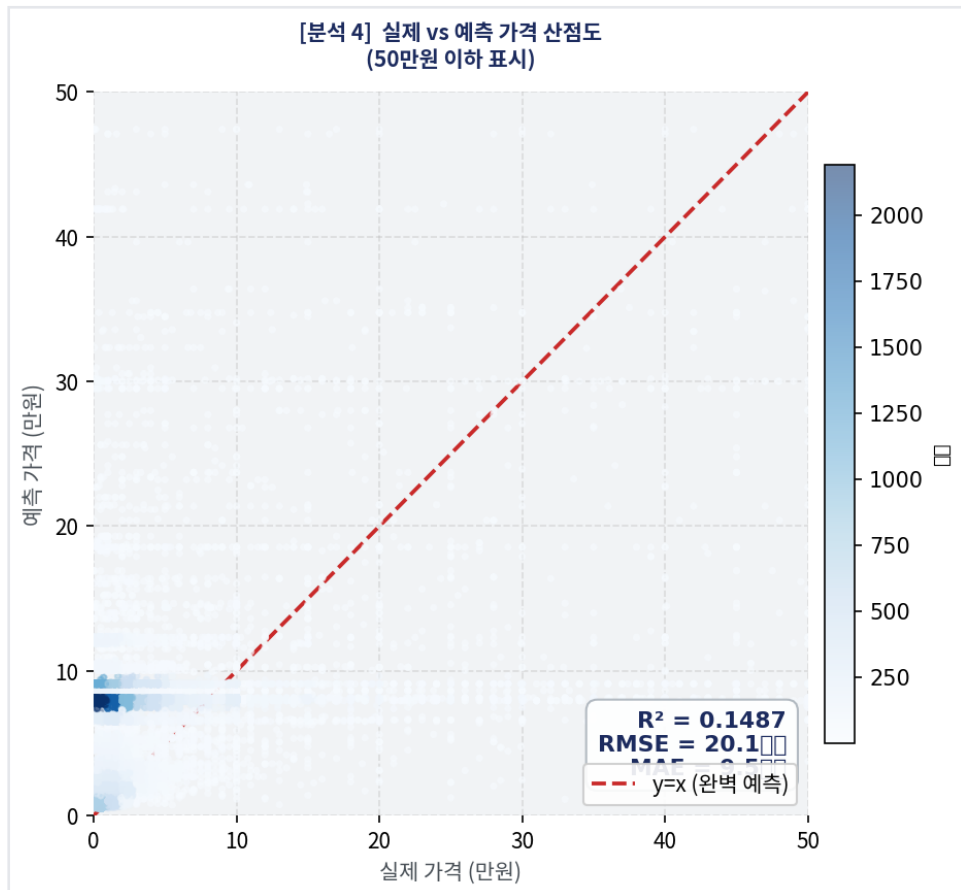
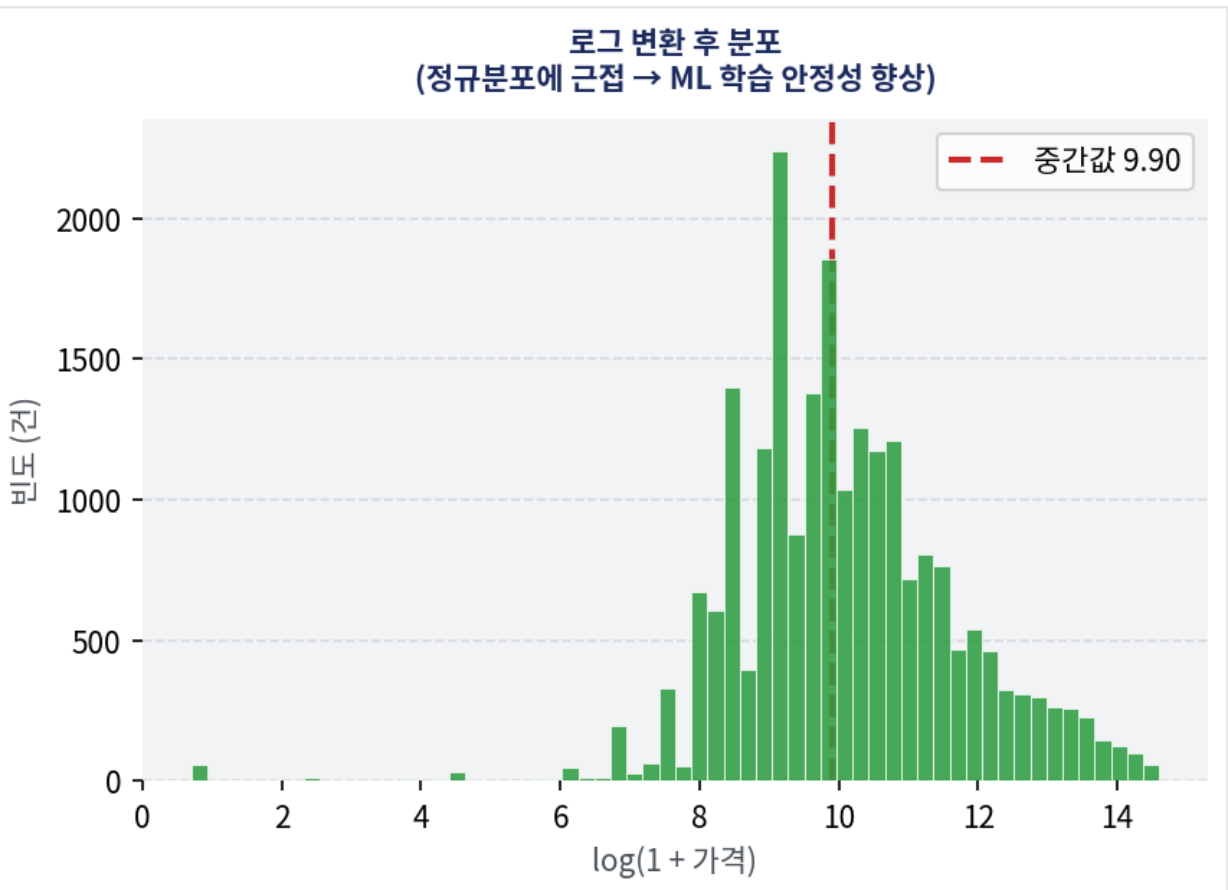
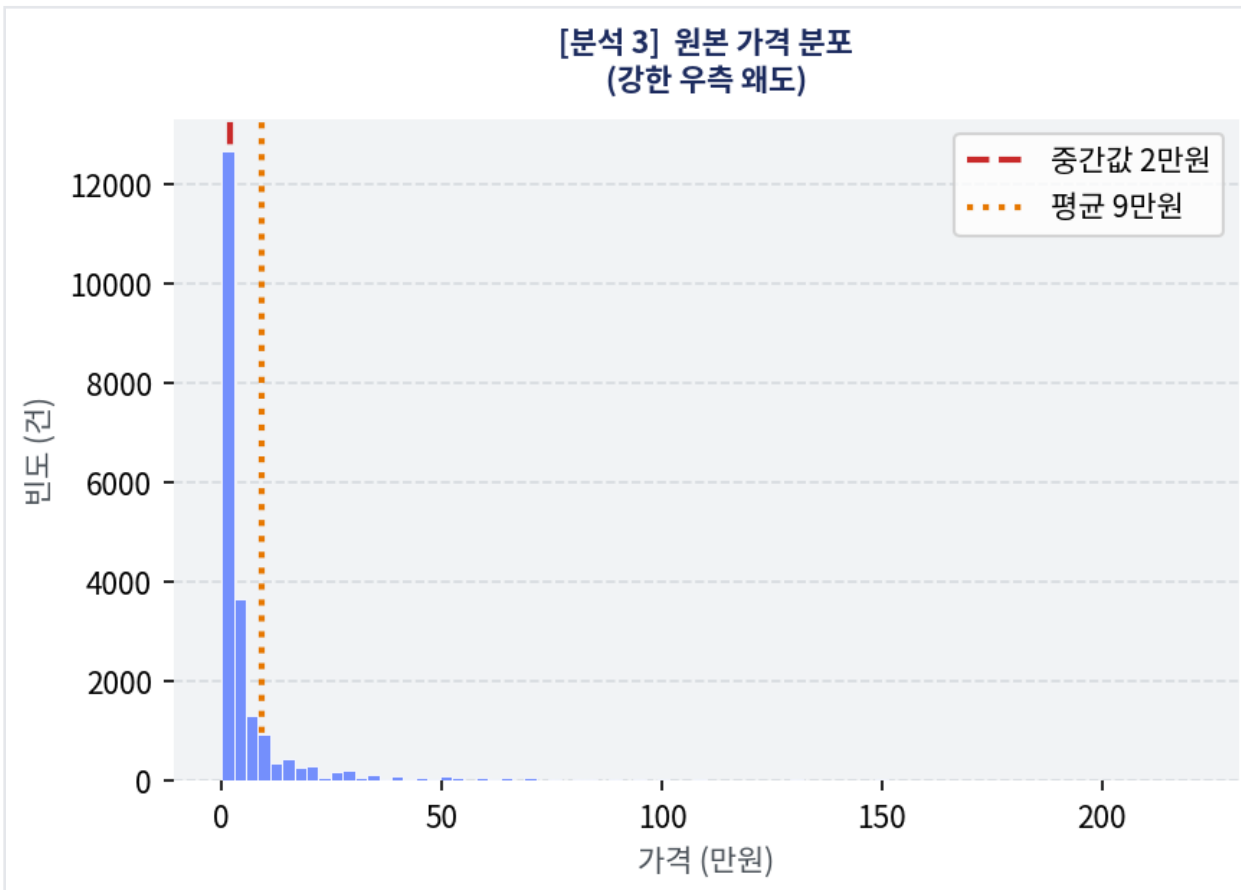
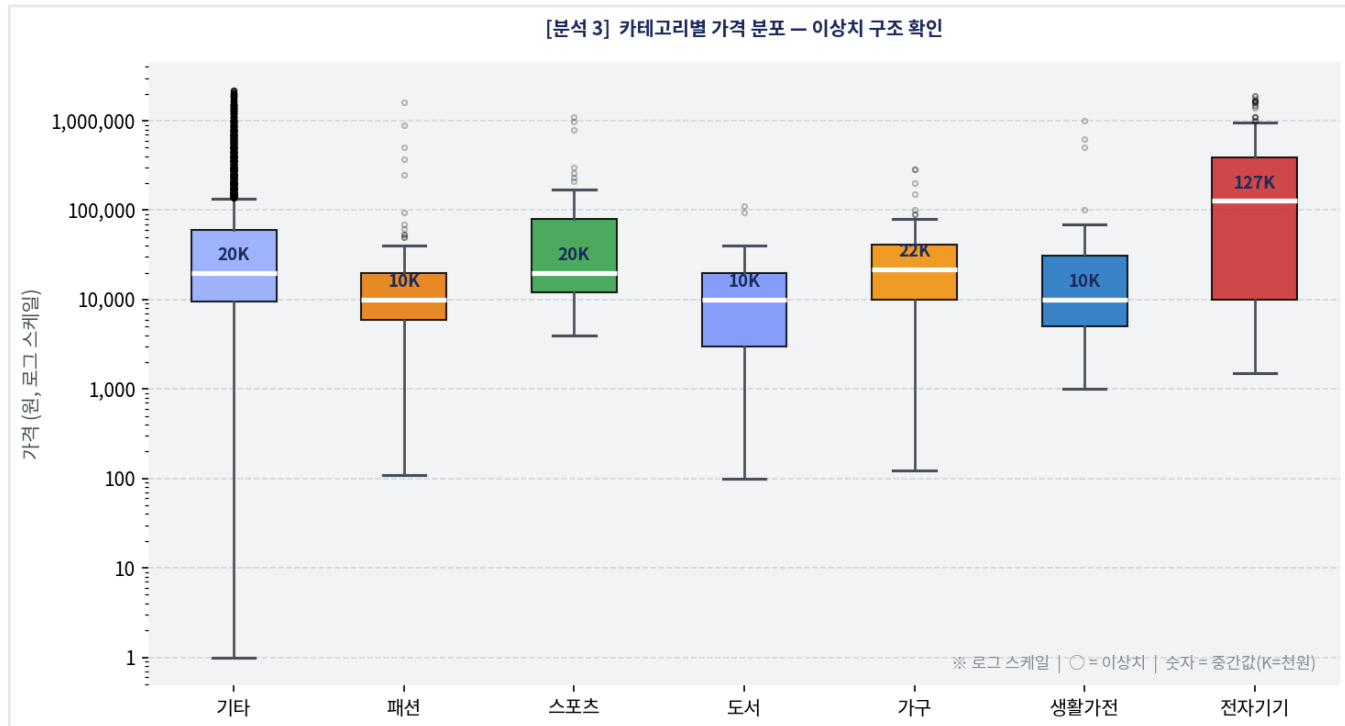
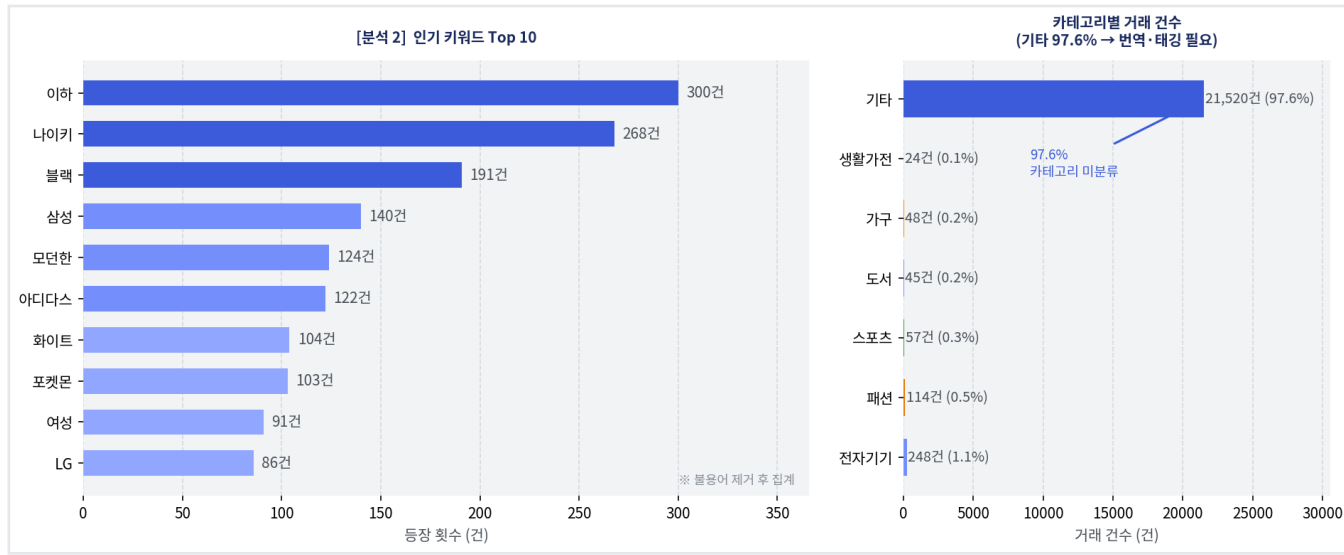
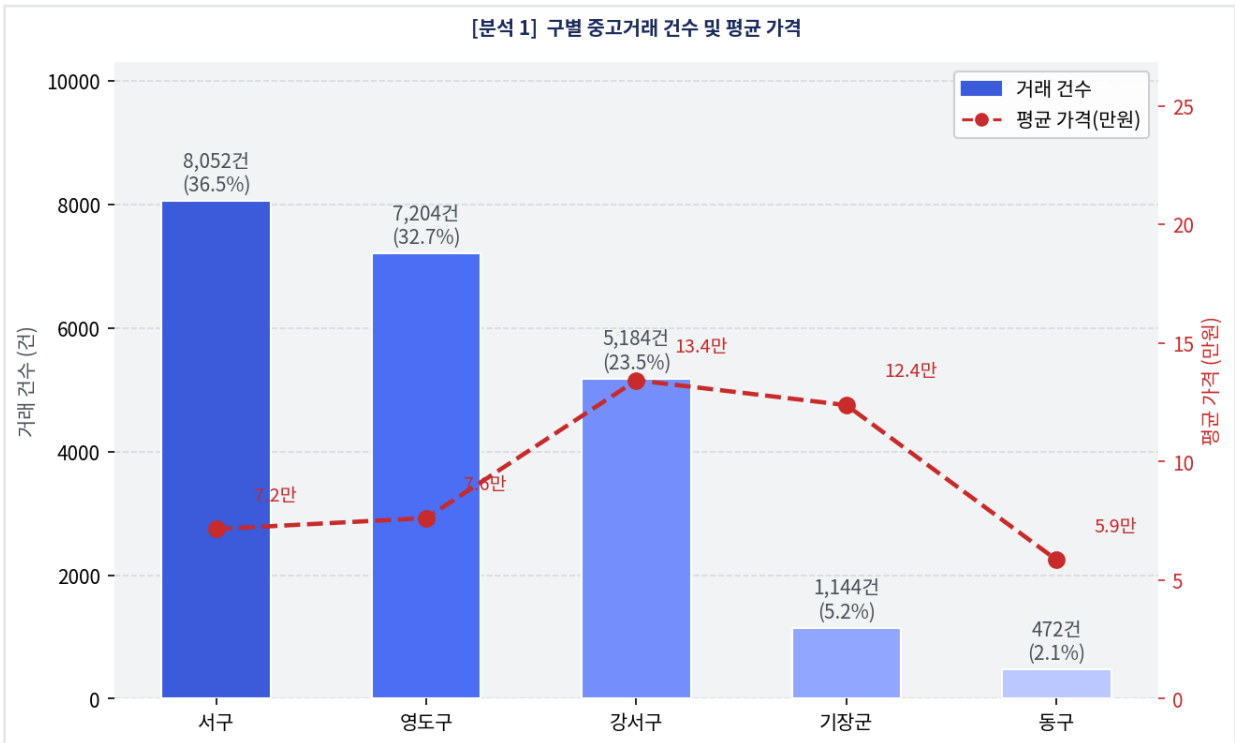
속성	값
크롤링 대상	당근마켓 (부산광역시 100개 동네)
원본 수집량	29,213건
최종 데이터	22,056행 × 9컬럼
결측치	0건 (원본 기준)
이상치 처리	상위 1% 아웃lier (2,200,000원 초과 제거)
수집 지역	서구 · 영도구 · 강서구 · 기장군 · 동구

※ 상세설명 · 조회수 · 관심수는 미포함 · 향후 추가 수집 시 모델 성능 개선 여지가 있다.

03

탐색적 데이터 분석 (EDA)

당근마켓 실제 수집 데이터(22,056건) 기반 분석 결과 · Matplotlib 산출물

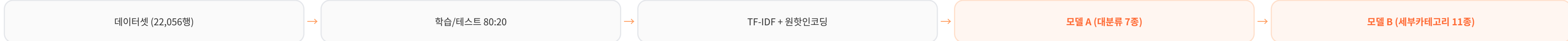


서구 · 영도구 · 강서구 3개 구가 전체 거래의 92.7%를 차지하며, 강서구 · 기장군은 거래량은 적지만 평균 가격이 높다. 카테고리별 97.6%가 판매자 미입력으로 '기타'로 분류되어 LLM 기반 자동 태깅의 필요성을 보여준다.

04

머신러닝 모델

scikit-learn 기반 RandomForestRegressor 학습 완료 · 재목 TF-IDF + 카테고리 원핫인코딩 + 제품길이를 특징으로 사용



성능 비교 (테스트셋 기준)

모델	RMSE	MAE	R ²
모델 A · 대분류 7종 (n_estimators=200)	202,848	98,167	0.074
모델 B · 세부카테고리 11종 (n_estimators=300)	201,317	97,020	0.088

세부카테고리 반영으로 R²가 소폭 개선(0.074→0.088)되었으나 예측력은 여전히 낮다. 고가 상품군(예: 태블릿) 학습 데이터 부족으로 과소 예측 우위가 있어, 정밀 가격 예측보다 '저렴 · 적정 · 비쌌' 3단계 판정 기능에 집중하는 방향으로 전환했다.

기술 스택: pandas, scikit-learn, beautifulsoup4, selenium, langchain-openai, streamlit, plotly

05

AI 번역 에이전트

LangChain + OpenAI(gpt-4o-mini) 연동 구현 완료 · 상품 정보 → 가격 판정 → 다국어 설명 생성까지 자동화

실제 테스트 사례 (삼성 갤럭시탭 · 영어 / 일본어 / 중국어 응답 생성 확인)

입력 · 상품 정보
삼성 갤럭시탭 S9 Ultra · 전자기기 · 판매가 550,000원 · 데이터 기준 참고가 612,000원

실제 가격 판정 로직 (judge_price)

LLM 출력 · 영어
1. Product Summary: Samsung Galaxy Tab S9 Ultra. 2. Price Check: Priced below the reference price - good deal. 3. Things to Check: battery health, accessories included.

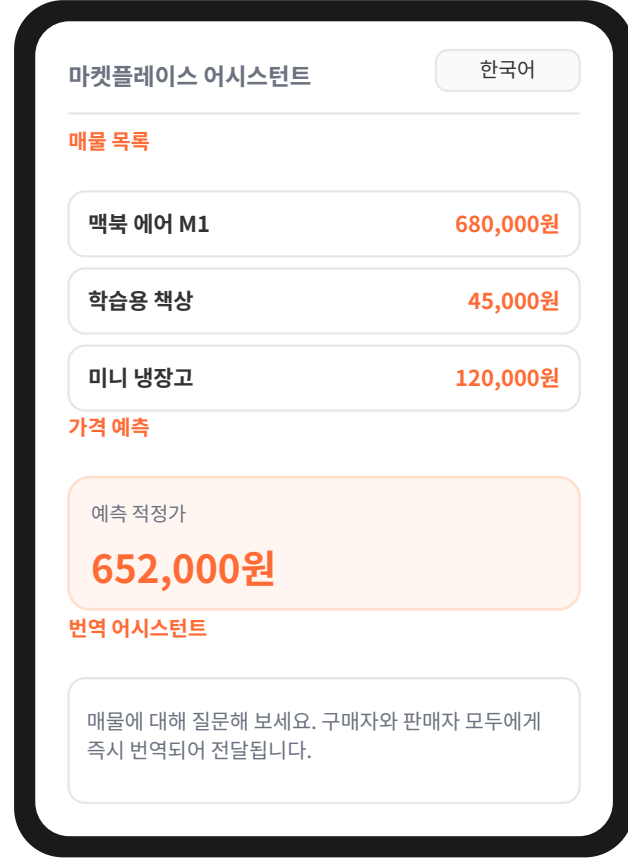
LLM 출력 · 일본어
1. 商品概要: サムスン ギャラクシータブ S9 Ultra. 2. 価格チェック: 参考価格より安く、お得です。3. 購入前確認: バッテリー状態、付属品。

LLM 출력 · 중국어
1. 商品简介: 三星 Galaxy Tab S9 Ultra. 2. 价格检查: 低于参考价格, 性价比高。3. 购买前确认: 电池状况、配件是否齐全。

06

대시보드

Streamlit 기반 대시보드 구축 · 아래는 목표 화면 예시



07

결과 및 인사이트

당근마켓 실제 데이터(22,056건) 분석 및 모델링 결과

92.7%

서구 · 영도구 · 강서구 3개 구에 거래가 집중 · 유학생 밀집지역의 소액 생활용품 거래 패턴을 보여준다.

기타 97.6%

판매자 미입력으로 대부분 매물이 미분류 상태 - LLM 자동 태깅의 필요성을 데이터로 확인했다.

R² 0.074→0.088

세부카테고리 반영으로 예측 성능이 개선됐지만 여전히 낮아, 3단계 가격 판정 방식으로 전환했다.

3개 언어 실증

gpt-4o-mini로 영어 · 일본어 · 중국어 상품 설명 생성을 실제 테스트로 확인했다.

08

고찰 및 한계점

현재 한계점	향후 개선 방향
카테고리 분류 한계 판매자가 카테고리를 직접 입력하지 않아 규칙 기반 분류 결과 97.6%가 '기타'로 남는다.	LLM 기반 자동 태깅 '기타'로 분류된 매물을 LLM으로 재분류해 카테고리 정확도를 높인다.
낮은 가격 예측력 R ² 가 0.07~0.09 수준으로 낮고, 태블릿 등 고가 상품군은 학습 데이터 부족으로 과소 예측될 수 있다.	학습 데이터 확장 상세설명 · 조회수 · 채팅수 등을 추가 수집해 예측 성능을 개선한다.
부가 데이터 부재 상세설명 · 조회수 · 관심수 등이 없어 재목 · 카테고리 기반 예측이 어렵다.	Streamlit 앱 구현 링크 입력 → 가격 판정 → 다국어 설명까지 이어지는 실제 대시보드를 구축한다.