김정훈

ML Engineer / AI Researcher

bandlagom0927@gmail.com
010-9559-9471 · 서울
portfolio.jhkim3ispark.work
github.com/thuygom
토스증권 AI Engineer (WM)

소개

자연어 질문에 필요한 정보를 찾고 검증된 실행 결과로 돌려주는 LLM 앱을 개발하고 있습니다. CRM 스키마 검색과 SQL 검증, 사용자별 접근 제어를 구현했고, 검색 근거와 음성 작업 상태를 서비스에서 확인할 수 있게 했습니다.

경력·학력

이즈파크 R&D Center · AI Researcher 2025.08–현재

일학습병행 인턴 2025.08–2026.02 / 정규직 2026.03–현재

TieMing의 AI 개발 2인 중 1인으로 명함 OCR/KIE, 검색, 실시간 STT, CRM Text-to-SQL과 대화형 음성 기능을 담당했습니다.

AI Summit Seoul 2026 공개 시연·Android 비공개 베타·Play Store 등록 준비 중

portfolio-tieming.jhkim3ispark.work

한성대학교 컴퓨터공학부·빅데이터 전공·공학사 2019.03–2026.02

GPA 3.95/4.5 · IRIS Lab 학부연구원 2024–2026 · 논문 5편, 제1저자 4편

주요 프로젝트

CRM Text-to-SQL · 사내 조회와 권한 제어 2026.06–현재

- DB의 약어·누락된 관계로 스키마 선택이 어려워 설명을 보완했습니다. Qdrant·Elasticsearch 검색 순위를 결합해 질문별 후보를 좁혔습니다.
- SQLGlot으로 단일 SELECT/CTE·허용 스키마·LIMIT를 검사하고, READ ONLY transaction과 timeout으로 실행 범위를 제한했습니다.
- 사내 영업 데이터의 사용자별 조회 범위를 제한하도록 PostgreSQL RLS 기반 행 접근 제어를 설계·검증했습니다.

portfolio-t2s.jhkim3ispark.work

iScribe·실시간 STT · 음성 처리 2026.03–현재

- AMI 회의 4건(경계 허용오차 0초·겹침 발화 포함)에서 **가중 DER 28.72→26.92%**인 Community-1을 선정했습니다.
- 실시간 STT의 5화자 연결음성 1건(enroll·loose 평가)에서 동기화·중복 전사를 수정해 **WER 62.77→34.06%**, DER 48.58→30.23%로 줄였습니다.
- 실시간 세션·장시간 job을 분리하고 PostgreSQL 원장으로 재시작 복구를 구현했습니다. GPU advisory lock·OIDC로 작업 진입과 tenant 접근을 제어했습니다.

iscribe.jhkim3ispark.work

체험 계정: demo / 비밀번호: demo

모델 학습·평가

GomOCR ·명함 OCR/KIE 후학습

2026.04–현재

- 검수 명함 2,906건을 train 2,034 / validation 581 / fixed test 291로 나누고 ID·이미지 SHA-256 교차 중복을 점검했습니다.
- Qianfan-OCR의 vision tower를 고정하고 language model·projector를 NF4 QLoRA로 후학습했습니다. RTX 4090 2장 DDP·BF16·effective batch 16을 적용했습니다.
- 후학습·출력 보정 후 기존 모델 대비 **필드 값 정확도 30.2→65.3%**, **목록 micro-F1 54.5→80.2%**를 기록했습니다(고정 test 291건).
- JSON 실패만 재생성해 같은 adapter의 **parse 오류를 27→1건**으로 줄였습니다. 조건은 validation에서 정했고 test는 이전 관측 세트입니다. 주소 F1·record exact의 소폭 회귀도 확인했습니다.

portfolio-gomocr.jhkim3ispark.work

Text-to-SQL ·별도 모델 학습

- Distil-Qwen3-4B를 NF4 QLoRA로 후학습했습니다(학습 2,172건). 동일 템플릿 분포의 **합성 검증 241건에서 SQL token 완전 일치 85.1%**를 기록했습니다.
- 별도 실제 100질문은 **실행 결과 완전 일치 45%**, pipeline 완료 95/100, 평균 value-Jaccard 0.767이었습니다. 학습 adapter와 CRM 서비스 추론 경로를 분리했습니다.

Web Search ·근거 수집

- 검색 API의 요약문과 실제 페이지 본문을 구분하고, 검색·크롤링 결과를 출처와 수집 상태를 포함한 공통 형식으로 정리했습니다.

개발·운영

- vLLM·FastAPI 추론 API와 Docker·cloudflared 배포를 구성했습니다. Prometheus·Grafana에서 서비스·컨테이너 상태와 GPU 메모리·온도·전력을 함께 관측했습니다.
- NIPA 고성능 컴퓨팅 지원사업의 **A100 2장**으로 LLM을 학습하고 있습니다. Weights & Biases로 실험을 관리하고 Hugging Face 비공개 저장소에 가중치를 보관합니다.

논문·특허

MRO ·제1저자—Multimodal Routing Optimization via Neural Architecture Search of Fusion Paths. IJIBC 17(3), 144–154, 2025.08 ·KCI 등재.

그 외—IN-MRO(KSC 2025.12), AURA(KIPS 2025.05), SCOPE(AICOMPS 2024)는 제1저자, SCOPE-X(KIPS 2025.11)는 공동저자입니다.

특허 출원·심사청구—데이터베이스 설명 자동 생성 및 환각 억제 기능을 갖는 Text-to-SQL 전처리 방법. 10-2026-0119188 ·2026.06.30 ·발명자 김정훈.

수상·자격·병역

수상—한성공학경진대회 대상(2023.09), 한성 SW페스티벌 최우수상(2024.11), ACK 2025 은상(2025.11), KSC 2025 장려상(2026.01), AI 스마트워크톤 은상(2026).

자격—SW개발 L5(2026), 정보처리산업기사·리눅스마스터 2급(2022).

병역—공군 정보체계 병 21개월·전문하사(2021.12–2022.12) 복무. Linux 서버·정보체계 관리.