FURIOSA

김정훈

bandlagom0927@gmail.com

ML Engineer / AI Researcher

010-9559-9471 · 서울

portfolio.jhkim3ispark.work

github.com/thuygom

지원 직무: Machine Learning Engineer (Agentic AI)

소개

AI 모델의 성능 개선부터 제품 적용까지 주도하는 ML Engineer입니다. 멀티모달 연구에서 쌓은 실험 역량을 바탕으로 TieMing의 주력 AI 기능을 개발하고 있습니다. 모델 학습부터 추론 API 구현과 서비스 배포까지 직접 맡아, 연구 아이디어를 실제 사용 가능한 기능으로 완성해 왔습니다.

경력

이즈파크 R&D Center · AI Researcher 2025.08–현재

일학습병행 인턴 2025.08–2026.02 / 정규직 2026.03–현재

TieMing의 주력 AI 기능인 명함 정보 추출·자연어 데이터 조회·회의 음성 처리를 담당합니다.

AI Summit Seoul 2026 공개 시연

대한민국 공군·정보체계 특기 전문하사 2021.12–2022.12 · 1년

Linux 서버를 관리하며 쉘 스크립트로 운영 작업을 처리하고 로그를 분석했습니다. 복무 중 정보처리산업기사와 리눅스마스터 2급을 취득했습니다.

학력

한성대학교 컴퓨터공학부·공학사 2019.03–2026.02

평점 3.95/4.5 · IRIS Lab 학부연구원 2024–2026 · 논문 5편 중 제1저자 4편

데이터 조회 AI · 학습과 실행 시스템 설계

CRM Text-to-SQL · 자연어 기반 영업 데이터 조회 2026.06–현재

Text-to-SQL 연구·특허 준비 2025.12 시작

- 사내 영업 데이터를 자연어로 조회하는 AI의 검색·SQL 생성 흐름을 설계했습니다. 잘못된 테이블 선택의 원인이던 DB 약어 설명과 누락된 테이블 관계를 보완하고, 의미·키워드 검색을 결합해 필요한 테이블을 선별했습니다.
- 교사 모델이 생성한 SQL 데이터와 LoRA 후학습으로 **내부 CRM 64문항의 실행 결과 일치율을 26.6%에서 76.6%로 50%p 높였습니다**. 행 순서·숫자·날짜 표현을 정규화해 기준 결과와 비교했습니다.
- 서비스 연동 전 실제 집계 결과까지 검증했습니다. 정상 실행된 SQL이 의도와 다른 날짜 열을 기준으로 집계하는 문제를 추적해, 잘못된 열 선택을 유발하던 불필요한 스키마 정보를 줄이고 실행 예제를 수정했습니다.
- **SQL 생성과 사용자별 데이터 접근 제어를 분리해 설계**했습니다. 복잡한 조인에서도 조회 권한이 적용되도록 PostgreSQL **행 단위 접근 제어(Row Level Security, RLS)**와 읽기 전용 실행을 적용·검증하고, 권한 회수·복원 시나리오를 점검했습니다.

모델 성능·서비스 처리 시간 개선

명함 정보 추출·모델 학습과 업무 적용 2026.04–현재

- 명함에서 이름·회사·직책을 항목별로 추출하는 **핵심 정보 추출(Key Information Extraction, KIE) 모델을 학습**했습니다. 글자 인식과 항목 배정 오류를 나눠 평가하고, 교사 모델의 지식을 작은 학생 모델에 전달하는 **지식 증류(Knowledge Distillation)와 LoRA 미세 조정**을 적용했습니다.
- 반복 생성으로 응답 형식이 깨지는 원인을 분석하고, 정상 결과는 보존한 채 형식 검사를 통과하지 못한 출력만 다시 생성하도록 구현했습니다.
- 업무에서 명함 **약 5,000장을 처리하며 핵심 7개 항목의 정보 추출 정확도 94%를 달성**했습니다. 별도 고난도 평가에서는 후학습과 출력 보정으로 항목 값 정확도를 **30.2%에서 65.3%로 높였습니다**.

회의 음성 서비스·전사 품질과 처리 시간 개선 2026.03–현재

- **장시간 녹음의 비동기 처리·복구 구조를 설계·구현**했습니다. 요청 제한 시간 때문에 결과를 받지 못하던 처리 방식을 바꾸고, 진행 상태와 중간 결과를 저장해 재시작 후에도 작업을 복구하도록 했습니다.
- 발화 구간 탐지(VAD)와 배치 처리를 적용했습니다. 기존 순차 처리와 비교해 **30분 녹음 파일 한 건의 서버 처리 시간을 122.6초에서 77.4초로 36.9% 단축**했습니다.
- 음성 인식·화자 분리 모델을 같은 조건에서 비교·선정해 제품에 통합했습니다. 내부 한국어 회의 기준 **단어 오류율(WER) 약 19%, 화자 분리 오류율(DER) 약 23%**를 달성했습니다.
- 영업·제조 전문 용어와 사용자가 추가한 표현을 용어 사전에 반영했습니다. Honcho 기반 사용자별 메모리에 개인 용어와 회의 맥락을 구분해 저장하고, 다음 음성 처리에 활용하도록 연동했습니다.

추론 서비스·검색 시스템 구축

- 모델을 API로 제공하기 위해 vLLM·FastAPI 기반 추론 서버와 Docker 배포 환경을 구축했습니다. Prometheus·Grafana로 서비스·컨테이너·GPU 상태를 함께 관측해 장애 원인을 추적할 수 있도록 했습니다.
- 후속 AI 기능이 검색 근거를 다시 검증할 수 있도록 원문·출처·수집 시점을 보존하는 검색 결과 구조를 설계했습니다. 수집 실패와 정상 결과를 구분하고 API 응답 형식을 통일했습니다.

연구·멀티모달 모델 설계

한성대학교 IRIS Lab ·멀티모달·자연어처리 연구 2024–2026

- 댓글 분석의 한계를 연구 과제로 발전시켜 영상·음성을 결합한 멀티모달 연구로 확장했습니다. 모델 구현·비교 실험·논문 작성을 수행해 **제1저자 논문 4편을 포함한 총 5편의 논문을 발표**했습니다.
- **입력에 따라 정보 결합 경로를 선택하는 모델을 설계**했습니다. 신경망 구조 탐색과 정책 학습으로 결합 방식을 탐색하고, 정확도와 연산량을 함께 비교해 모델 구조를 평가했습니다.

논문

정보 결합 경로의 정확도·연산량 연구—Multimodal Routing Optimization via Neural Architecture Search of Fusion Paths

IJIBC 17(3), 144–154 ·KCI 등재·2025.08 ·제1저자

입력에 따른 멀티모달 경로 선택—암시적 NAS 정책학습 기반 멀티모달 경로 최적화

KSC 2025.12 ·제1저자·장려상(2026.01 수상)

콘텐츠와 댓글을 함께 분석한 사용자 반응 연구—멀티모달 인플루언서 지지도 추정을 위한 Attention 기반 사용자 반응 분석

KIPS 2025.05 ·제1저자

창작자 보호를 위한 정서 지표—창작자 보호를 위한 SNS 정서 리스크 지표화

KIPS 2025.11 ·공동저자·은상

댓글 감정·주제를 이용한 인기 예측—A Transformer-Based System for Predicting Influencer Popularity via Comment Sentiment and Topic Analysis

AICOMPS 2024 ·제1저자

특허 출원

데이터베이스 설명 자동 생성 및 환각 억제 기능을 갖는 Text-to-SQL 전처리 방법

DB 설명과 관계 정보를 보완해 잘못된 SQL 생성을 줄이는 전처리 방법을 설계했습니다.

출원·심사청구·10-2026-0119188 ·2026.06.30 ·발명자 김정훈

주요 기술

모델 학습—Python, PyTorch, Transformers, PEFT, 지식 증류, LoRA/QLoRA

추론·서비스—vLLM, FastAPI, WebSocket, PostgreSQL, Elasticsearch, Qdrant

운영—Linux, Docker, Prometheus, Grafana

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 특기 병으로 21개월 복무.