

김정훈

ML Engineer / AI Researcher

portfolio.jhkim3ispark.work
토스 AI Engineer (Model)

bandlagom0927@gmail.com

010-9559-9471 · 서울

github.com/thuygom



소개

명함에서 글자를 잘못 읽는 오류와 이름·회사 항목을 바꾸어 넣는 오류를 나눠 개선했습니다. 영업 데이터 질문 조회에서는 잘못된 표 선택, 실제 결과, 사용자별 접근 권한을 따로 검증했습니다. 모델 점수뿐 아니라 서비스에서 실패하는 조건까지 평가하고 수정합니다. 제가 올해 가장 자주 활용한 학습 기법은 LoRA 미세조정과 교사 모델의 지식을 학생 모델에 전달하는 지식 증류(Knowledge Distillation, KD)입니다.

경력·학력

이즈파크 R&D Center · AI Researcher

2025.08-현재

일학습병행 인턴 2025.08-2026.02 / 정규직 2026.03-현재

AI 관계 관리 제품 개발 2인 팀에서 명함 정보 추출, 사람·회사 검색, 회의 음성 전사와 영업 데이터 조회를 맡았습니다.

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

평점 3.95/4.5 · IRIS Lab 학부연구원 2024-2026 · 논문 5편, 제1저자 4편

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

모델 학습·평가

명함 정보 추출·정확도 개선

2026.04-현재

- 실제 업무 OCR 약 5,000장 규모에서 핵심 7개 항목 정보 추출 정확도 94%를 확인했습니다. 별도 후학습 평가의 항목 값 정확도는 30.2%에서 65.3%로 높였습니다.
- 기존 모델 전체 대신 적은 수의 추가 가중치만 학습하는 LoRA 미세조정과 교사 모델 지식을 학생 모델로 전달하는 지식 증류(KD)를 명함 추출 모델 학습에 적용했습니다. 문자 인식(OCR)과 핵심 정보 추출(Key Information Extraction, KIE)을 나눠 평가하며 KIE 이론과 평가 방법을 익혔습니다.

portfolio-gomocr.jhkim3ispark.work

자연어 데이터 조회 모델의 별도 실험

- 별도 학습 모델은 같은 질문 틀의 합성 검증 241건에서 조회문 완전 일치율 85.1%를 기록했습니다. 같은 형식의 검증이므로 실제 질문 정확도로 보지 않았습니다.
- 실제 질문 100건에서는 처리 경로 완료 95건, 기준 조회 결과와 완전 일치 45건이었습니다. 이 모델은 운영 중인 조회 서비스와 분리했고 복잡한 여러 표 연결 질문의 실패를 따로 확인했습니다.

서비스 적용

영업 데이터 질문 조회·사용자별 접근 제어

2026.06-현재

- 데이터 약어와 빠진 관계 설명 때문에 질문에 필요한 표를 찾기 어려웠습니다. 설명과 관계를 보완해 조회문 생성 전에 후보를 좁혔습니다.
- 생성 조회문에서 사용자 조건이 빠져도 다른 직원의 자료가 나오지 않도록 PostgreSQL 행 단위 접근 제어(Row Level Security, RLS)를 설계했습니다. 조회문 검사와 읽기 전용 실행을 함께 적용했습니다.
- 격리된 검증 환경에서 두 사용자로 같은 질문을 실행해 타인 자료가 0건인지 확인했고, 권한 회수 뒤에 도 결과에서 사라지는지 검증했습니다. 운영 연결은 담당자와 협업했습니다.

portfolio-t2s.jhkim3ispark.work

회의 음성 전사·작업 복구

2026.03-현재

- 사내 회의록에서 영업·제조 용어를 반영한 사전을 만들고, 새 표현을 사용 중 누적했습니다. Honcho 기반 사용자별 메모리로 개인 단어와 회의 맥락을 분리해 다음 작업에 활용했습니다.
- 화자 1-10명의 고난도 한국어 음성 10건을 같은 조건으로 비교해 iScribe를 탑재했습니다. 단어 오류율(WER) 29.40%·화자 구분 오류율(DER) 33.25%였고, 실사용 기준은 각각 약 19.4%·23.3%였습니다.
- 실시간 세션과 장시간 작업을 나누고 처리 단계·중간 결과를 저장해 종료·재연결 실패 뒤에도 중단 지점부터 복구했습니다.

iscribe.jhkim3ispark.work

학습 재개·GPU 자원 관리

- 중단 뒤에도 이어 학습할 수 있도록 모델·최적화 상태를 저장했습니다. GPU 온도와 전력을 30초마다 기록하고 82도에 도달하면 저장 후 학습을 멈췄습니다.

개발·운영

- 모델 API와 컨테이너 배포를 구성하고, 서비스·서버·GPU 상태를 한 대시보드에서 확인해 장애 원인을 좁혔습니다.
- 정보통신산업진흥원 고성능 컴퓨팅 지원으로 제공받은 NVIDIA A100 GPU 두 장에서 언어 모델을 학습하고 있습니다.

논문·특허

멀티모달 정보 결합 경로 최적화·제1저자—Multimodal Routing Optimization via Neural Architecture Search of Fusion Paths. IJIBC 17(3), 144-154, 2025.08 · KCI 등재.

그 외—입력에 맞춘 정보 결합 연구(KSC 2025.12), 댓글·영상 반응 기반 지지도 추정(KIPS 2025.05), 댓글 기반 인기 예측(AICOMPS 2024)은 제1저자입니다. 창작자 보호 정서 지표 연구(KIPS 2025.11)는 공동저자입니다.

특허 출원·심사청구—자연어 질문을 데이터베이스 조회문으로 바꾸기 전에 설명과 관계를 보완하는 전처리 방법. 10-2026-0119188 · 2026.06.30 · 발명자 김정훈.

수상·자격

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).