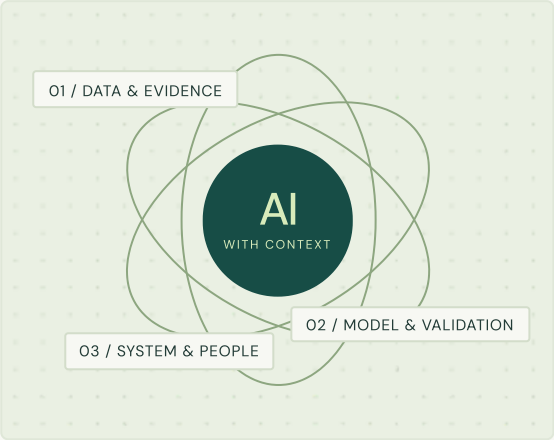
AI PORTFOLIO / 오서연

현장의 문제를 이해하고, 실제로 쓰이는 AI를 만듭니다.

도메인의 맥락을 기술로 풀어내는 오서연입니다.

FOCUS LLM 서비스 개발과 운영 (LLMOps) | AI 에이전트 개발 | AX / 업무 AI 전환

은행 여신 업무의 AI 기능 기획과 사내 배포, 이커머스 상품 데이터 정비, 비개발 동료와 함께한 평가 설계까지. 현업의 문제를 과제로 정의하고, 데이터로 효과를 확인해 왔습니다.



EMAIL osy.seoyeon5@gmail.com
WEB sohds.github.io
GITHUB github.com/sohds
LINKEDIN linkedin.com/in/seoyeon-oh-soh6012

한국산업은행 / 업무 개선
KDB GPT 배포
종합의견 기능을 제안·설계, 현업 검토 후 사내 배포

정보경영연구실 / 데이터 정비
결측률 68% → 17%
전체 상품 속성 기준, 결과물은 협력사 서비스에 반영

미래에셋증권 AI Festival / 평가 설계
Hit@5 93.8%
팀이 만든 438문항 중 채점 419문항, 내부 검색 평가

01 / EXPERIENCE

현업의 문제를 AI 과제로

업무를 이해하고, 과제로 정의하고,
효과를 확인하기까지.

한국산업은행
Korea Development Bank

2026.06 — 2026.08
한국산업은행
AI디지털전략부 인턴
AI개발팀 | AI데이터팀

LLM Serving Metrics
Knowledge Engineering

01 여신 업무 AI Assistant 기획과 사내 배포

공문 초안, 산업분석 요약, 문체 변환 등 업무별 시스템 프롬프트를 설계했습니다. 직접 제안한 여신승인신청서 종합의견 생성 기능은 RM과 CO의 검토 관점을 나눠 설계했고, 지점 실무진의 검토를 거쳐 사내 LLM 서비스(KDB GPT)에 배포되었고, 회장 보고에 팀 사례로 포함되었습니다.

System Prompt | Domain Analysis | Expert Review | 서비스 배포

02 LLM 사용량을 업무 목적별로 보는 운영 지표 설계

서버 과부하 사례를 계기로, 기존 vLLM 지표를 목적별로 보는 Grafana 대시보드를 만들어 제안했습니다.

AS-IS
전체·서버 단위 집계라, 과다 호출이 어느 용도에서 생겼는지 가리기 어려움

→

TO-BE
새 수집 없이 목적별로 재집계해, 처리량과 TTFT·TPOT·E2E p95로 튄 용도를 좁힘

vLLM | Grafana | Prometheus | PromQL

데모(가상 데이터) sohds.github.io/demo/vllm-dashboard

03 전결권 규정 구조화 PoC: 조건과 결재권자의 관계 보존

원문을 프롬프트에 넣으면 표 안의 표가 무너져 조건과 결재권자의 대응이 끊겼습니다. 규정을 규칙 단위로 나누고 원문 위치를 붙이는 LLM Wiki를 제안해, 가상 규정으로 모델 4개를 비교 검증했고 팀 추진 방향으로 채택되었습니다.

LLM Evaluation | Markdown / YAML | Schema Design

KDB / 문제를 어떻게 정의하고 검증했나요?

과제 발굴: 담당자가 판단에 집중하도록

참고 문서에서 여신승인신청서의 종합의견 생성 과제를 발굴하고, 여신 교재 6권과 작성 가이드라인을 학습했습니다. RM과 CO의 관점 차이를 반영해 역할별 Assistant를 설계했으며, 원문 보존, 범위 제한, 적합한 Assistant 안내를 프롬프트에 반영했습니다.

현업 검토로 확인한 현장 관행

테스트케이스로 프롬프트를 조정한 뒤 지점 실무진의 검토를 받았습니다. 이 과정에서 문서상 작성 방식과 현장 관행의 차이를 확인했고, 서술 형식과 재무제표 기준이 보완된 최종본이 배포되었습니다.

운영 지표를 읽는 기준

Counter와 Gauge의 특성에 맞춰 집계를 구분하고, 평균에 가려지는 느린 요청을 보기 위해 지연의 p95를 시각화했습니다. 트래픽 증가와 사용자 체감 성능을 함께 확인하는 운영 관점을 정리했습니다.

데이터로 모델 선택: 결함을 세어 비교

같은 가상 규정으로 모델별 깨진 링크, 필수 섹션 누락, 원본에 없는 내용 생성을 비교했습니다. 결함이 없던 모델로 만든 규칙 9개는 원문 줄 번호와 모두 일치했습니다. 개정 실험에서는 규칙 ID 재사용과 삭제된 예외의 이력 추적 문제를 확인했습니다.

조건부 결론까지 보고

직급 중심과 업무 중심의 엔티티 구성을 5개 업무 영역에 걸쳐 비교했습니다. 업무 중심 구조의 장점은 세분화 기준이 명확할 때 유지된다는 조건을 도출하고, 형식 검증과 의미 검증을 분리하는 방안을 제안했습니다.

LLM Wiki는 인턴 개인 PoC의 실험과 비교 결과입니다. 구조는 팀 추진 방향으로 채택되었으나 이후 도입은 보류되었으며, 정량적 성능 개선 수치는 주장하지 않습니다.

이화여자대학교 EWHA WOMANS UNIVERSITY

2024.04 — 2025.08

정보경영연구실

Information Management Lab
학부연구생 및 외부연구원
산학 협력 과제
2024년 서울여자대학교,
2025년 이화여자대학교 소속

Conversational Recommendation
RAG
Image Generation
서비스 반영

01 2024.04 — 2024.10 / 4인 팀

고객의 실시간 쇼핑 상태를 반영하는 가상 휴먼 대화 엔진 (AI Sales Assistant)

상품 상세페이지의 글과 이미지에서 누락된 속성을 추출하고 정규화해, 전체 상품 속성의 결측률을 68%에서 17%로 낮췄습니다. QA와 추천 모듈, 통합 API 설계를 맡아 상품명, 의도, 속성을 분리해 검색하는 RAG 흐름을 구성했고, 직접 작성한 100문항을 연구실 구성원 5명이 정오로 판정한 정성평가에서 QA 정답률을 개선 전보다 30%p 높였습니다. 추천에는 SQL로 추출한 사용자 행동 데이터를 반영했습니다.

Data Quality · SQL · LangChain · FAISS · FastAPI

02 2024.08 — 2024.10 / 단독 개발

온라인 셀러를 위한 AI 글쓰기 서비스

QA 답변의 말투 실험에서 출발해 온라인 셀러의 상품 설명문을 생성하는 서비스를 단독 개발했습니다. 말투별 세부 지시문과 예시 문장을 설계해 생성 문체의 편차를 줄였으며, 상용 서비스에 적용되어 상품 설명문 작성을 자동화했습니다.

Prompt Engineering · Brand Voice · 상용 서비스 적용

03 2025.05 — 2025.08 / 4인 팀

패션 이커머스 상세페이지 자동 생성 서비스

의류 이미지 생성 트랙을 단독으로 맡았습니다. 첫 과제에서 품질을 정성평가로만 설명한 한계를 의식해 SSIM, LPIPS, CLIPScore 같은 정량 지표를 검토하고, 생성 결과물을 직접 비교하며 LLM 평가용 루브릭을 설계했습니다.

Runway Gen-4 · VLM · Prompt Engineering · Rubric Design

산학 협력 과제의 결과물은 협력사의 실제 서비스에 반영되었습니다. 패션 상품의 색상, 핏, 소재 등 누락 속성은 TTA 속성 체계에 맞춰 정규화했고, 일반 대화, QA, 추천 모듈은 의도별 라우팅을 갖춘 통합 API로 연결했습니다.

02 / SELECTED PROJECTS

금융 분야 팀원과 함께 검증한 프로젝트

전체 Repository
github.com/sohds

FROM DOCUMENTS
TO A DEPLOYED AGENT

DOCUMENTS
문서와 펀드 데이터

↓

HYBRID RAG
검색, 계산, 근거 검증

↓

CLOUD API
Docker / Nginx / NCP

MIRAE ASSET 미래에셋증권

2026.07 — 2026.09

미래에셋증권 AI Festival / 3인 팀 프로젝트

연금 AI Agent / 연금꿀사

제공 문서만 근거로 연금 제도, 세제, 펀드 비교 질문에 답하는 상담 Agent입니다. PDF, DOCX, XLSX, PPTX 문서 58건을 형식별로 파싱하고, PDF는 PyMuPDF와 Docling을 비교해 제목과 표가 보존되는 결과를 골랐습니다. 금융 분야 팀원 2명과 협업하며 기술 설계와 구현, 서버 배포를 총괄했습니다.

Python | HyperCLOVA X | Hybrid RAG | Docker | Nginx | NCP

58 → 358
원본 문서 → 검색 청크

100
구조화한 펀드 데이터

93.8%
내부 근거 문서 Hit@5

답이 92위로 밀린 이유를 청킹에서 찾음
중도인출 질문의 답이 다른 청크로 갈려 BM25 92위에 있었습니다. 모델은 분할 앵커만 반환하고 코드가 원문을 추출하도록 바꿔, 재검증에서 1위가 되었습니다. BM25와 임베딩 검색은 RRF로 결합했습니다.

현업 관점의 검수를 데이터로 연결
원문과 분할 경계를 나란히 보고 판정을 남기는 검수 화면을 만들었습니다. 두 팀원은 판정 369건 중 35건에 문제를 표시했고, ISA 재가입 유의사항이 떨어져 검색된다는 지적을 받아 분할 방식을 고쳤습니다.

클라우드 서버 구축과 배포
Docker로 패키징한 API 서버를 NCP에 배포하고 Nginx를 역방향 프록시로 구성해, 재부팅 후 복구까지 확인했습니다.

평가 범위
팀 평가셋 438문항 중 채점 대상 419문항에서 근거 문서 Hit@5 93.8%, 정상 질문 오차단 0건. 계산 모듈 최종 수정 전 E2E 요청 459건의 실패 0건을 확인했습니다.

원문 16건 비교
설명 없이 사라진 글자 21.9%를, 사유를 남긴 제외 구간으로 전환

숫자 검증 결함
0.7억을 7억으로 읽던 단위 변환을 고치고 금액 표기 9종을 회귀 검사로 고정

92위→1위는 공모전 후 재검증한 질문 1건의 BM25 순위로, 인덱스 구성도 달라 청킹만의 효과가 아닙니다. Hit@5는 상위 5개 검색 결과의 근거 문서 포함률이며 답변 정확도나 공식 심사 점수가 아닙니다.

FIN:NECT

2026.03 — 2026.07

BITAmin 팀 프로젝트 / PM / FIN:NECT 챌린지 서울 지역예선 참가

멀티에이전트 교차 토론 기반 데일리 주식 시황 브리핑 시스템

Bull과 Bear 에이전트의 교차 토론을 투자 브리핑으로 전달합니다. 거시경제, 섹터, 시장 심리 데이터로 2라운드 토론을 구성하며, 자동 주문 기능이 없는 브리핑 도구입니다.

Python | LLM Agents | Streamlit

github.com/sohds/multi-agent-trading

BITAmin

2026.01 — 2026.03

BITAmin 프로젝트 / PM

한국형 액티브 ETF 종목 선정 및 백테스팅 시스템 개발

외국인과 기관 순매수 기반 종목 선정과 백테스팅 시스템입니다. 리밸런싱 주기, 수급 신호, 비중 방식, 가격 기준을 바꿔 비교하는 실험을 구성하고, Streamlit 대시보드에서 NAV, 종목 구성, 최대 낙폭과 샤프 비율 등 성과 지표를 확인할 수 있습니다.

pandas | Backtesting | Plotly

github.com/sohds/Ko-ActiveETF

서울여자대학교 SEOUL WOMEN'S UNIVERSITY

2023.07 — 2023.08

Programming GURU2 Android 해커톤 / 4인 팀 / PM / 장려상

VegeVision: 농산물 신선도 분류 Android 앱

PM 겸 팀 대표로 Kotlin 앱 개발, 딥러닝 모델과 Firebase 연동을 담당했습니다. MobileNetV2 전이학습 모델을 TensorFlow Lite로 변환해 앱에서 추론하도록 연결하고, 팀원들의 화면 작업을 통합했습니다.

Kotlin | TensorFlow Lite | Firebase

github.com/sohds/VegeVisionApp



2023.07 — 2023.10

자율주행 인공지능 알고리즘 개발 챌린지 / 팀 아이들 / 3등 장려상

주행 영상 객체 탐지 모델 개발

팀 전원이 영상 인식 경험이 없는 상태에서 한 달간의 학습 계획표를 제안했습니다. 주행 이미지 10만 장에서 1만 장을 선별하고, PyTorch로 YOLO 기반 차량과 보행자 탐지 모델을 구현했습니다.

PyTorch | YOLO | Object Detection

한국교통안전공단 주최, 국토교통부 주관

03 / PUBLICATIONS

연구로 남긴 질문들

Google Scholar

2025.02 — HCI KOREA / pp. 1187–1192

Grad-CAM++와 Activation Map을 활용한 한국인 여성 중심 딥페이크 탐지 특징 분석

김규리, 김세인, **오서연**, 조유라, 최다예, 최덕우, 최예림

2024.11 — 대한전자공학회 추계학술대회 / pp. 665–669

환자 메타데이터를 활용한 호흡 및 음성 소리의 대조 학습을 통한 SARS-CoV-2 양음성 및 중증도 진단

오서연(제1저자), 김다영, 김예린

04 / ABOUT ME

문제를 구체화하고, 끝까지 구현합니다.

현업의 문제를 과제로 정의합니다

한국산업은행에서는 여신 교재 6권으로 업무 구조를 익혀 담당자의 판단을 돕는 기능을 골랐고, 지점 실무진의 검토를 거쳐 배포했습니다.

현업이 검수에 참여하게 만듭니다

연금 AI Agent에서는 금융 분야 팀원들이 원문과 검색 청크를 직접 대조하고 의견을 남길 수 있는 검수 화면과 질문 양식을 만들었습니다.

데이터로 효과를 확인합니다

구현 전에 평가 기준부터 정하고, 결측률, 검색률, 결함 건수처럼 확인할 수 있는 지표로 개선 효과를 설명합니다.

함께 사용하는 도구들

Languages & Data
Python, SQL, Kotlin, pandas, scikit-learn, Firebase, Knowledge Graph

LLM & Retrieval
LangChain, FAISS, BM25, RRF, Hybrid RAG, HyperCLOVA X, Prompt Engineering

Generation & Evaluation
Runway Gen-4, VLM, Rubric Design, SSIM, LPIPS, CLIPScore

Deep Learning
PyTorch, TensorFlow Lite, MobileNetV2, YOLO, Grad-CAM++

Deployment & Monitoring
FastAPI, Docker, Nginx, NCP, vLLM, Grafana, Prometheus

활동

2026.01 — 2026.07 — **연합 빅데이터분석 동아리 비타민(BITAmin)**
금융 데이터와 AI 프로젝트 활동

2024.04 — 2025.08 — **이화여자대학교 정보경영연구실**
학부연구생 및 외부연구원

2023.08 — 2025.02 — **인공지능 프로젝트 팀 다이브(daiv)**
AI 연구와 프로젝트 활동. 우수 동아리원 상 2회, 우수 프로젝트 상

수상

2023 — **3등 장려상** 자율주행 인공지능 알고리즘 개발 챌린지
팀 아이들

2023 — **장려상** GURU2 Android 해커톤
팀 베지비전

2022 — **장려상** GURU1 Python/HTML 해커톤
개인

자격

2025 — **SQLD** SQL개발자
한국데이터산업진흥원

2025 — **ADsP** 데이터분석준전문가
한국데이터산업진흥원