AI TECH LEAD · MLOPS SYSTEM ARCHITECT

비즈니스 문제를 아키텍처와 AI 시스템으로 해결하는 엔지니어

AWS 아키텍처 설계부터 Multi-Agent 최적화까지, AI 관련 요구사항을 시스템으로 구현합니다.

문제 해결 사례 보기 | PDF 포트폴리오

FOCUS	DOMAIN	STACK
MLOps / Agent	AI · 제조 · 법무	FastAPI · AWS

01 — STRENGTH POINTS

한 줄로 맡길 수 있는 세 가지

End-to-End MLOps 인프라

Terraform · CloudFormation 기반 IaC로 AWS 인프라를 올리고, FastAPI 백엔드부터 딥러닝 모델 서빙까지 학습-평가-배포 루프를 하나의 시스템으로 설계합니다.

LangGraph Multi-Agent 도입

LangGraph · LangChain · CrewAI로 Agentic AI와 RAG를 설계합니다. LLM뿐 아니라 CV, Time Series, NLP 모델을 프로덕션 워크플로에 붙인 경험이 있습니다.

리더십 · 커뮤니케이션

기획·디자인 공백이 있어도 고객 핵심 요구를 아키텍처로 번역합니다. 이해관계자에게 구조를 설명하고 범위를 조율해, 팀이 실제로 배포할 수 있는 결정을 만듭니다.

02 — KEY PROJECTS

요구사항과 어떤 제약이 있었고, 시스템을 어떻게 바꿨는지

프로젝트별 핵심 문제와 해결 과정을 정리했습니다.

01 / 2026.04 — 2026.10

Production MLOps

글로벌 자동차 부품 제조사 MLOps

아이엔소프트 · Acoustic / Vision 모델 운용 시스템

AWS 위에서 Acoustic 파이프라인을 유지 보수하며, Vision 데이터 학습-배포 루프를 새로 얹는 과제였습니다. 현장 카메라 부터 SageMaker 학습 파이프라인까지 매끄럽게 연계하는 것에 중점을 두었습니다.

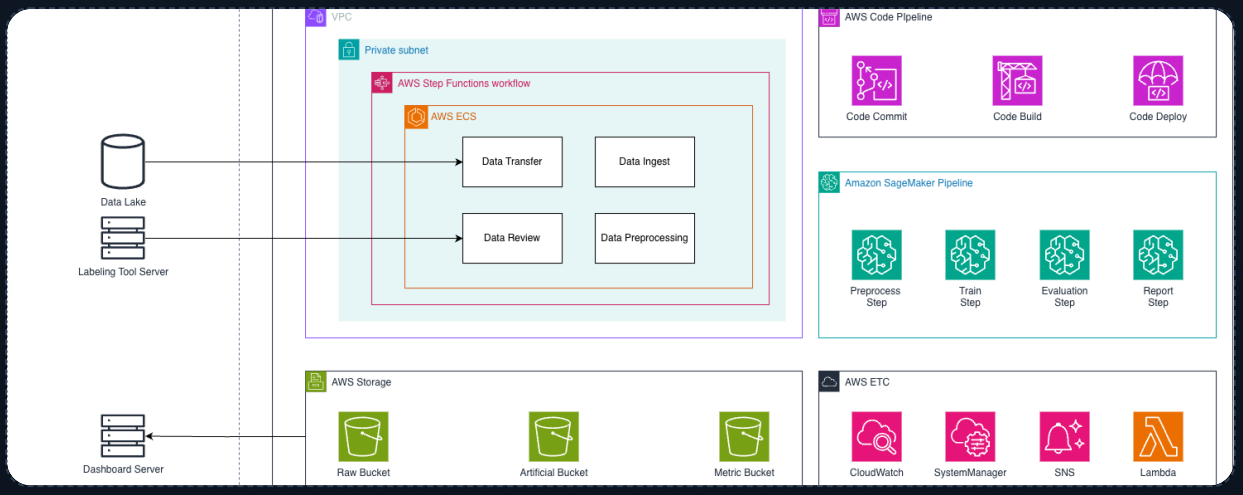
Python 3.11 · FastAPI · PyTorch · SMP · YOLO · SageMaker · ECS · Lambda · Step Functions · CloudWatch

주요 해결

현장 ROI ↔ 학습 루프 단절. SageMaker 지표를 대시보드 UI에 렌더링하고 현장 카메라 ROI를 학습 데이터와 연동해, 재학습이 현장 조건 위에서 돌도록 파이프라인을 닫았습니다.

Segmentation 전이학습 비용. 전이학습 전략을 고정하고 Loss, Early Stopping, AMP를 도입해 인스턴스 사용을 줄이면서 학습 효율을 올렸습니다.

운영 가시성. 수집-전처리 구간을 Step Functions 중심으로 묶고, 학습 및 평가 결과를 대시보드가 연동할 Restful API를 FastAPI로 제공했습니다.



02 / 2026.03 — 진행중

Multi-Agent Ops

iSODA cowork

아이엔소프트 · Agent 기반 인시던트 탐지·대응

장애를 챗봇에게 물어보는 시스템이 아니라, 탐지 Agent가 상황을 읽고 대응 Agent가 Runbook · Playbook을 실행하는 오케스트레이션이 필요했습니다.

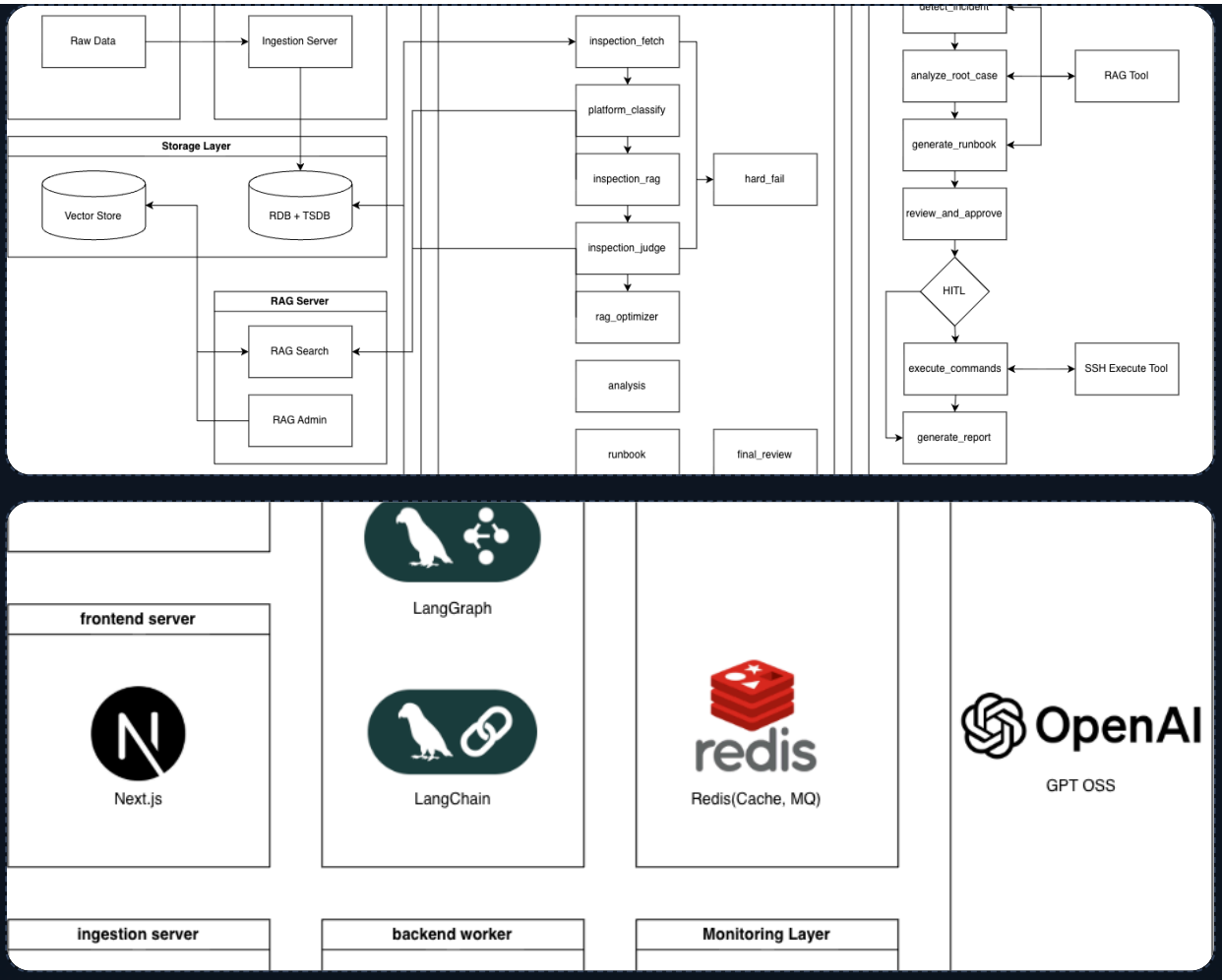
Python 3.11 FastAPI LangGraph LangChain vLLM Elasticsearch LangFuse Grafana Prometheus PostgreSQL Docker

주요 해결

선형 Agent의 한계. 일방향 질의응답은 인시던트 분기를 표현하지 못합니다. LangGraph 기반 Multi-Agent로 상황별 tool을 동적으로 호출하고 시나리오를 라우팅하도록 바꿨습니다.

PoC를 운영 언어로. 탐지 Agent를 검증할 데모 데이터와 시나리오를 먼저 설계해, 인프라(API · DB · 모니터링)가 실제 장애 흐름을 타도록 맞췄습니다.

관측 가능성. LangFuse와 Grafana · Prometheus를 같은 레이어에 두어, Agent 궤적과 시스템 메트릭을 함께 볼 수 있게 했습니다.



03 / 2025.07 — 2025.09

RAG · Agentic AI

SmartCLM

BSG 파트너스 · L사 법무팀 PoC · AWS Marketplace SaaS

계약 생성 · 검토 · 승인을 LLM 채팅으로 끝내는 제품이 아니었습니다. 문서를 파싱하고, 조항을 비교하고, 위험 조항을 남기는 작업 그래프가 필요했습니다.

Python 3.13 FastAPI SQLAlchemy LangChain LangGraph Bedrock CloudFormation ECS RDS PGVector Redis

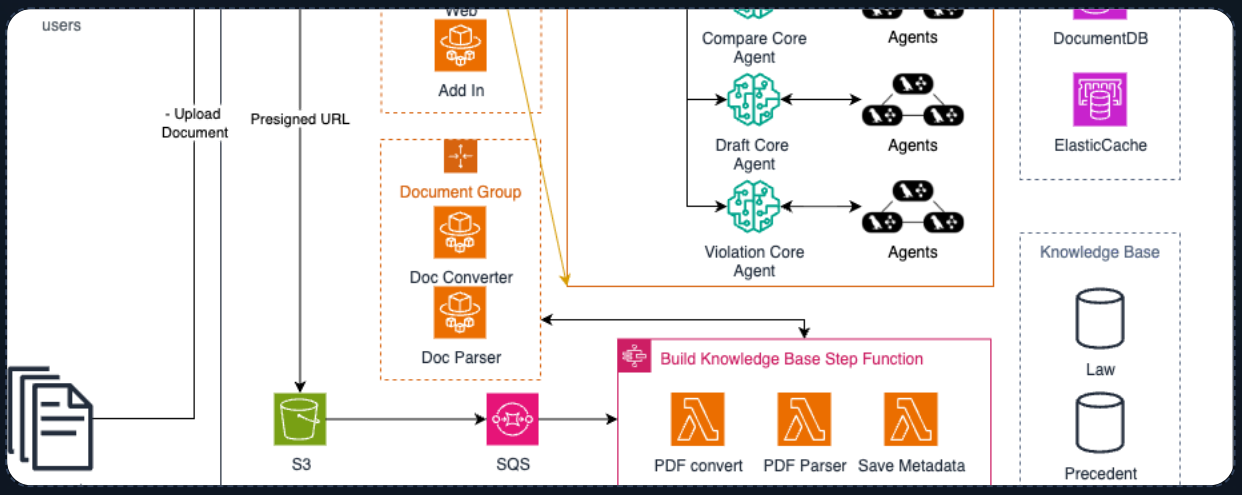
주요 해결

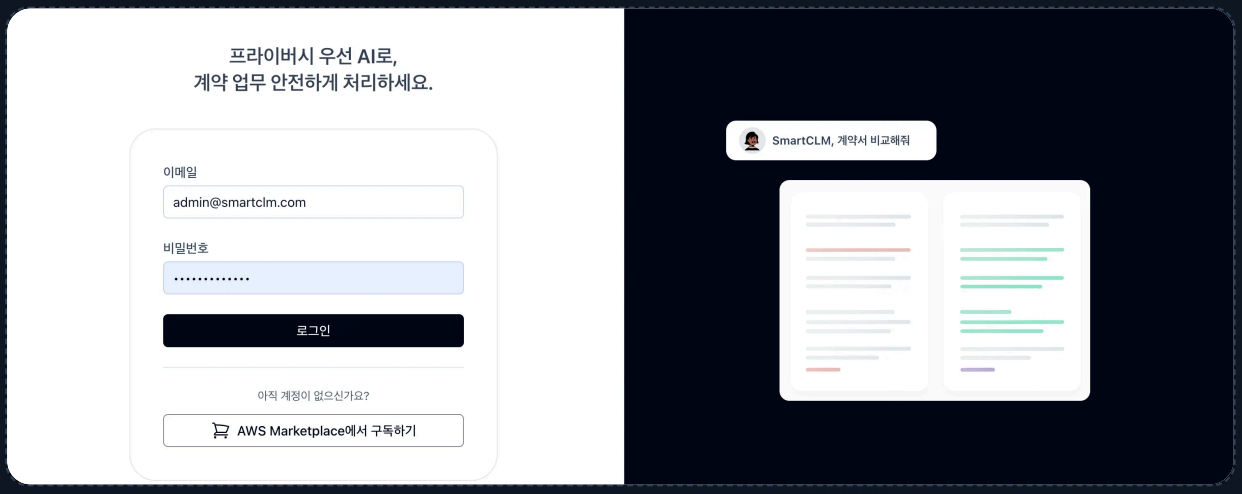
프롬프트 난립. LangChain으로 프롬프트와 Agent를 모듈화해, 계약 작업 유형이 늘어도 한 파일에 로직이 쌓이지 않게 만들었습니다.

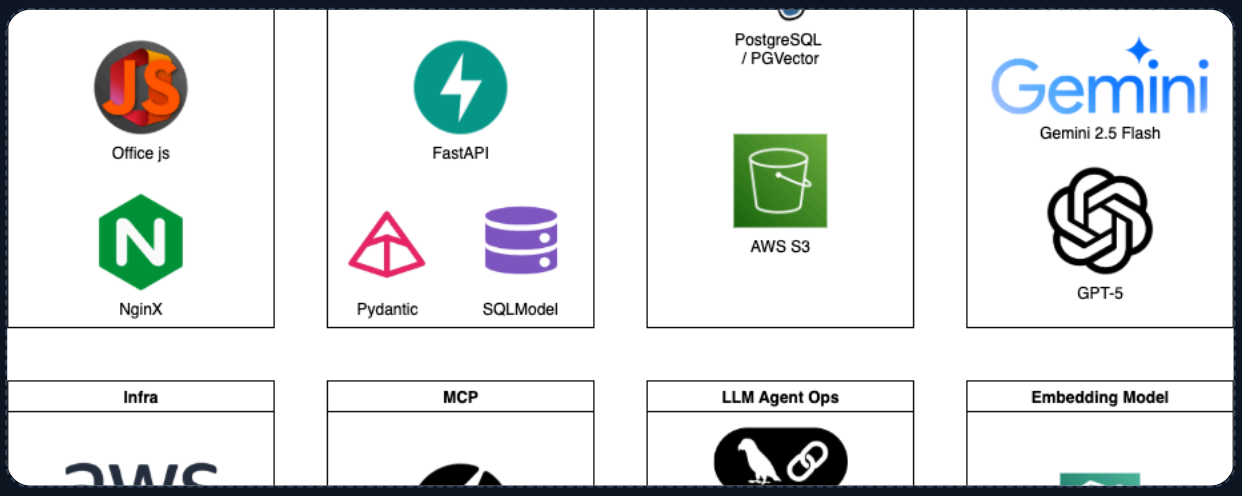
단일 Agent의 복잡도. Route / Summary / Compare / Draft / Violation Agent를 LangGraph로 분리해 검토·승인 같은 문서 작업을 그래프 경로로 타게 했습니다.

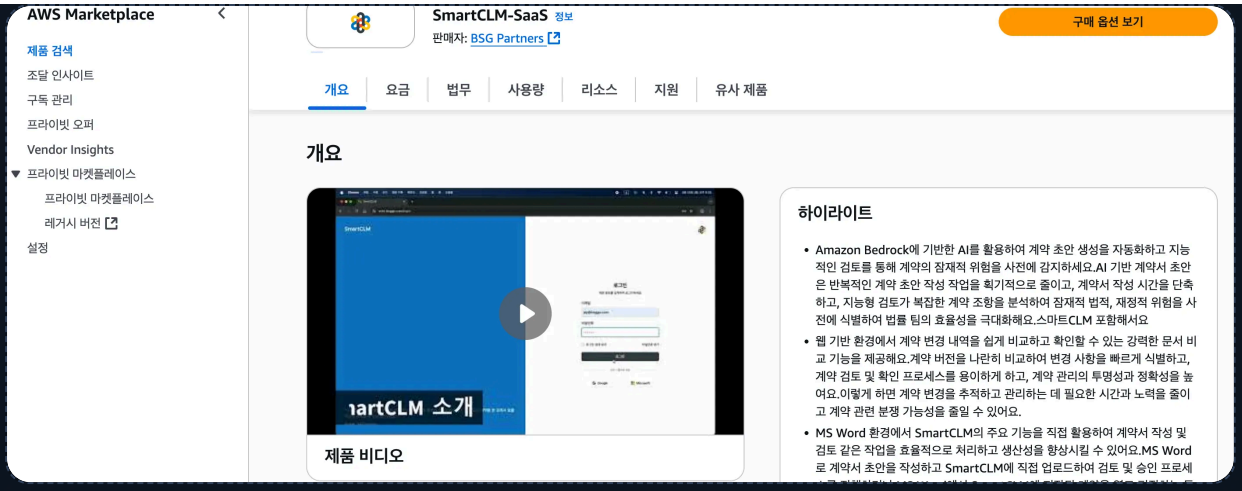
RAG 품질과 보안. 업로드 → PDF 변환 → 파싱 → Chunk/Embedding → PGVector · S3 적재 파이프라인을 두고, Agent는 DB 직접 연결 대신 API tool만 사용하도록 막았습니다.

제품화. AWS Marketplace entitlement · metering을 구현하고, L사 법무팀 PoC에서 정량·정성 지표를 통과했습니다.



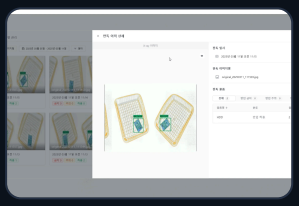






ADJACENT PRODUCTION WORK

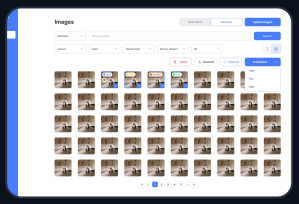
카드를 누르면 화면과 아키텍처, 해결 내용이 열립니다.



X-ray Detection

멀티스레드 + WebSocket으로 프레임 드롭 없는 실시간 Detection 스트림. OpenCV 캡처보드 해상도 유지.

자세히 보기 →



딥레이크

이미지 데이터셋 MLOps. Cognito RBAC, S3/SQS/Lambda 대용량 업로드, 업로드 응답 35% 개선.

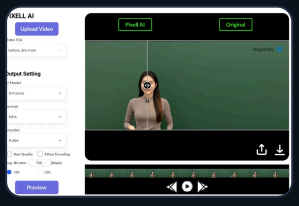
자세히 보기 →



스마트공장 이상탐지

RabbitMQ → WebSocket 센싱 파이프라인. 통계 조회 120ms → 38ms, 피크 전력 제어 효과.

자세히 보기 →



PMS · NPU Denoise

Furiosa NPU + Docker 모델 서빙, FFmpeg 인코딩. GPU/NPU 이원 추론 환경을 제품으로 납품.

자세히 보기 →

03 — SIDE PROJECT

실무 밖에서 기획부터 배포까지

Personal · Mobile

새김

AI 기반 성경 구절 타이핑 모바일 앱

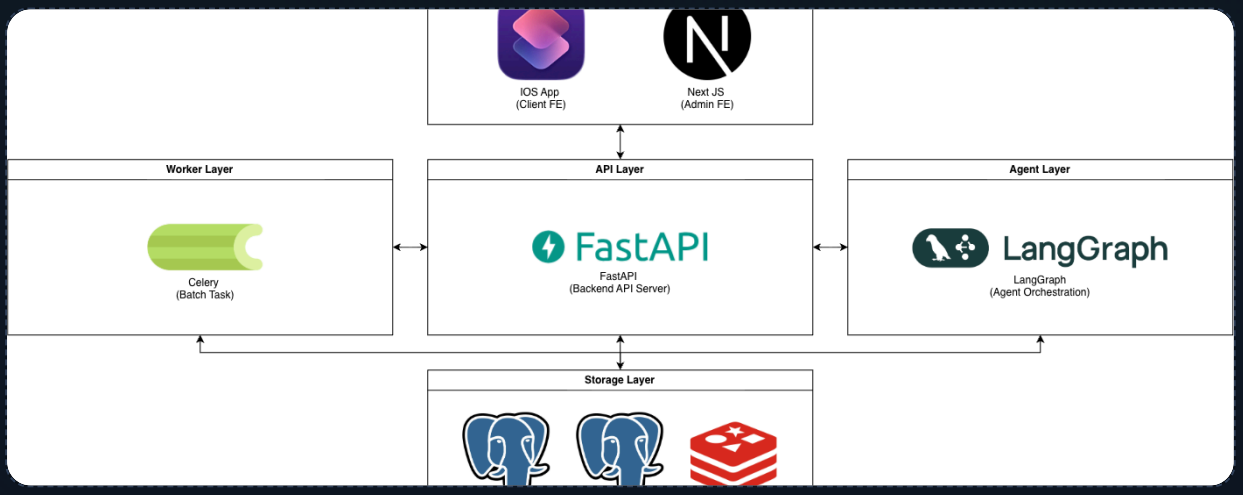
읽는 경험을 타이핑 루틴으로 바꿨습니다. 사용자가 남긴 속도와 패턴을 바탕으로 다음 구절을 고르는 추천 에이전트를 붙이고, 화면 흐름은 직접 설계했습니다.

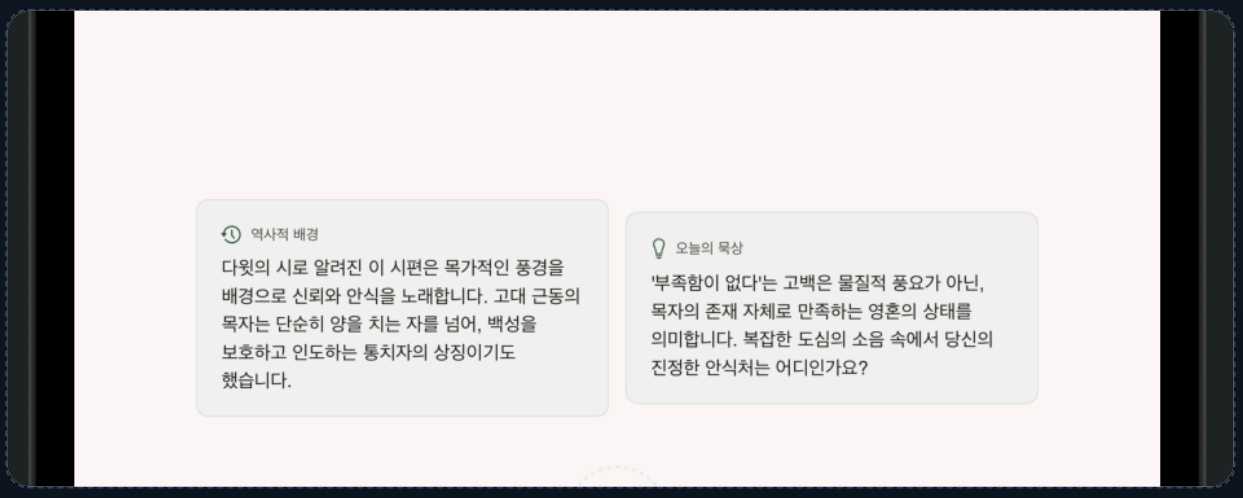
풀스택. React 클라이언트와 FastAPI 서버를 직접 설계·구현했습니다.

추천 에이전트. 사용자 맞춤형 AI 추천을 서비스 흐름에 연동했습니다.

UI/UX. 기획 공백 없이 화면 설계를 주도해 타이핑 경험의 밀도를 맞췄습니다.

React · FastAPI · AI Agent · UI/UX





최근 경력, 시스템이 커진 구간

아이엔소프트 · AX본부 AI시스템팀

2026.02 — 현재 · 프로

AWS ECS · Lambda · S3 · Step Functions · SageMaker 아키텍처와 LangGraph 기반 동적 Agent, LangFuse · Grafana · Prometheus 관측 스택을 설계합니다. 글로벌 자동차 부품사 MLOps와 사내 iSODA cowork이 이 구간의 본류입니다.

BSG 파트너스 · 클라우드본부 BD팀

2025.07 — 2025.09 · 과장 (계약직)

CloudFormation · CodePipeline 기반 CI/CD, Bedrock Agent, RAG 모듈화를 맡았습니다. SmartCLM을 AWS Marketplace SaaS로 올리고 L사 CLM 고도화 PoC를 끝냈습니다.

에이아이네이션 · 기업부설연구소

2023.01 — 2025.05 · 팀장 → 선임연구원

온프레미스 3-tier와 AWS 전환, Gitlab · Jenkins · Harbor CI/CD, PyTorch 모델 패키징까지 한 팀의 백엔드·인프라를 이끌었습니다. 딥레이크, X-ray Detection, 스마트공장, NPU 화질개선이 이 기간의 산출물입니다.

2022년 이전 경력

거민시스템 · 미라클솔루션 · 데이터/기획 프로젝트

펼치기

거민시스템 · 과장 2020.09 — 2022.12

정부과제 기획·PM, NLP/시계열/영상 모델, 데이터 파이프라인. 지하공동구 표준관리시스템, 디지털트윈 실증, 지역경제 빅데이터 플랫폼.

미라클솔루션 · 대리 2020.02 — 2020.08

네트워크 시계열 분석, 단국대 교육지식플랫폼 연계 모듈, TTA 인증 지원.

프리랜서 2019.10 — 2020.01

NLP · 시계열 · 고객 분석, API 기반 데이터 수집.

프로덕션에서 반복한 도구

Backend

FastAPI, Flask, SQLAlchemy, Redis, PostgreSQL, Kafka, WebSocket, Celery

AI / Data

PyTorch, LangGraph, LangChain, Bedrock, YOLO, Transformer, SageMaker, vLLM

Infra

AWS ECS/Lambda/S3, Terraform, CloudFormation, Docker, Kubernetes, Grafana

06 — TALKS

강의와 발표로 남긴 설명

아키텍처를 코드로만 두지 않고, 현장과 교실에서 같은 언어로 풀어 본 기록입니다.



2025 · AWS 센터필드

AWS AI Industry Week Demo Showcase

비에스지파트너스 SmartCLM 서비스를 시연하고, RAG·Agent 구조를 참관객에게 설명했습니다.



2023 · 한국공학대학교 탄소중립혁신센터

스마트에너지플랫폼 MC기술세미나

제조 현장 AI 특강. 센싱 데이터와 이상탐지를 스마트공장 맥락에서 풀었습니다.

2023 · 고려대 안암캠퍼스

심리학 기반 AI 이해와 활용 특강

AI 빅데이터 분석 프로세스를 비전공 청중 기준으로 재구성해 강의했습니다.

CONTACT

시스템 이야기를 이어가겠습니다

AI Tech Lead · MLOps 아키텍트 포지션의 문제 정의, 제약, 기대하는 운영 수준을 듣고 싶습니다.

veroroot@gmail.com

github.com/veroroot

© 2026 장진원. 경영학 전공 · ADsP · Coursera MLOps.