

여정훈

AI 엔지니어 · 풀스택 — LLM을 프로덕션에 신뢰할 수 있게 올립니다

aofhd0003@gmail.com · 포트폴리오 portfolio.eldsoft.com · GitHub github.com/YeoJeongHun1

2020년 개발을 시작해 7년차. 데이터 엔지니어링·백엔드를 거쳐 현재는 공공입찰 공고문 LLM 추출·AI 보완검증 플랫폼의 AI 파이프라인을 단독 설계·구축·운영하는 AI 엔지니어입니다. 프롬프트 세대 관리, 모델 실측 평가, 환각 방어, 비용 최적화까지 프로덕션 LLMOps 전 주기를 실무로 다루며, 1인 법인으로 AI 앱 2종 (ELANFIT·Conoti)을 스토어에 출시해 운영하고 있습니다.

핵심 역량 STRENGTHS

① LLM 프로덕션 엔지니어링

프롬프트 세대 관리(v1→v5, 실패 1건=규칙 1개 회귀셋) · 지연-트리거 병렬 race 풀백(fast path 99.2%) · 2-tier 모델 escalation · grounding 환각 게이트 · 검증 정확도 0.67→1.00

② AI 비용·품질의 정량 운영

모델 4종+ 실측 비교 선정(비용 8.68배 검증) · 명시적 캐시 -14.2% · 전처리 토큰 -74.2% · 책임분리로 검증 비용 -33.5% · 호출별 비용 전량 로깅

③ 데이터 파이프라인·인프라

Airflow ETL/ELT · RabbitMQ 멀티워커 · 레거시 DB 무중단 dual-write 이관(다운타임 ≤1분) · Kafka 스트리밍 · Docker/GCP/AWS·Terraform IaC

④ 제품을 끝까지 만드는 실행력

기획→설계→개발→출시→운영 1인 완주 (앱 2종 스토어 출시) · MCP 기반 AI 에이전트 하네스 구축 · 사내 AI 세미나 기획·강의

기술 스택 SKILLS

LLM (Gemini · OpenAI)

Prompt Engineering

RAG

MCP · AI 에이전트

LLMOps

Python

FastAPI

RabbitMQ

Apache Airflow

Kafka

MySQL

PostgreSQL

Redis

React

TypeScript

Vue 3

Flutter

Node.js

Spring

TensorFlow

Docker

GCP · Terraform

AWS

NGINX

경력 요약 CAREER

- **비드코칭연구소** 2025.04 – 현재 · AI 엔지니어
공공입찰 공고문 LLM 추출·AI 보완검증 플랫폼 — AI 파이프라인 설계·구축·운영 단독
- **주식회사 이엘디 (1인 법인 대표)** 2024.05 – 현재 · 겸직
ELANFIT(AI 피트니스, 2026.02 출시) · Conoti(AI 공지 자동화, 2026.06 출시) · 외주 2건
- **한국전력정보** 2025.03 – 2025.04
DR(수요반응) 자원 플랫폼 — FastAPI 백엔드 · React 웹
- **다임즈** 2021.09 – 2024.04 · 데이터 엔지니어
KT Dr.IP·IPaSDN(트래픽 예측 AI: LSTM/GRU/Prophet) · 보령/천안 신재생에너지 플랫폼 · KTOA
- **더안전한세상 · 누리엔지니어링** 2020.08 – 2021.05
보안관제 시스템(MFC) · 설계 연구

학력 국립한밭대학교 전자제어공학과 졸업 (2020.02)

자격 정보처리기사 · PC정비사 2급 · ITQ · 워드프로세서 1급

수상 프로보노 공모전 은상(2019, 전국) · RC페스티벌 금상(2018, 전국) 외



비드코칭연구소 · AI 엔지니어 — 설계·구축·운영 단독

2025.04 – 현재 · 운영 중

입찰공고 문서(HTML·HWP·PDF·이미지)에서 자격조건·금액 등 핵심 항목을 LLM으로 추출하고, 별도 AI 워커가 재검증하는 프로덕션 LLM 데이터 플랫폼. 설계→비용분석→구현→모델 선정→운영배포→고도화까지 LLMOps 전 주기를 단독 수행. (업무 자산 특성상 내부 세부는 비공개 — 아키텍처·기법·성과 수준으로 기술)

99.2%

기본 모델 fast path

0.67→1.00

검증 정확도(유사도)

-33.5%

검증 비용 (책임분리)

-74.2%

전처리 입력 토큰

v1→v5

프롬프트 세대

AI 엔지니어링 핵심

- **멀티 프로바이더 풀백·병렬 race** — 장애 로그 실측 분석(타임아웃·쿼터가 실패의 다수) 기반으로 "기본 모델 우선 + 지연 시에만 보조 모델 병렬 투입" 설계. 패자 응답도 백그라운드 수확·로깅.
- **모델 실측 평가·2-tier 운용** — 동일 프롬프트 1:1 비교(비용 8.68배 차 검증) 등 4종+ 평가로 운영 모델 선정. 평시 경량 모델, 근거 확인 실패 항목만 상위 모델 escalation.
- **AI 보완검증 워커** — "AI는 추출만, 판정은 워커" 책임분리 재설계로 비용 -33.5%·응답 -40.4%. 이슈코드 34종 등급 체계.
- **환각 방어 계층** — AI 값이 입력 원문에 없으면 환각 판정(grounding 게이트) + 1회 재추출 중재 + 별도 LLM을 정답지로 쓰는 배치 적대 감사.
- **프롬프트 세대 관리** — DB 기반 버전·모델 관리(무결성 해시), 실패 케이스 1건 = 규칙 1개 회귀셋. 검증 정확도 0.67→1.00.
- **비용 최적화·관측성** — 명시적 컨텍스트 캐시(-14.2%, 고정 프롬프트 단가 90%↓), 호출별 토큰·비용·풀백 경로 전량 로깅 + 비용 대시보드.

파이프라인·운영 자동화

- 비AI/AI 큐 분리 멀티워커(운영·개발·테스트 3중 환경), 레거시 DB 무중단 dual-write 이관(다운타임 ≤1분, Airflow 증분·역등)
- 정제 정규식 2,000여 개 수동 관리 → 차집합 기반 자동학습(제안→영향도→사람 승인, 유사 패턴 67→1 병합 실증)
- AI 변환 품질게이트(실측 분포 기반 임계) + 상위 모델 escalation, 동일 공고문 그룹당 AI 1회 호출(중복 ~3% 제거)
- AI 에이전트용 커스텀 MCP 하네스(읽기 전용·화이트리스트 도구 수집 중) · 작업이력 DB화 · 일일보고 자동화

스택 Python · FastAPI · RabbitMQ · Airflow · MySQL · Gemini/OpenAI · Docker · MCP · React 상세

portfolio.eldsoft.com/projects/procurement-ai-platform

대표 프로젝트 ② ELANFIT — AI 피트니스 코칭 앱 **LAUNCHED**



이엘디 · 1인 개발 (기획·모바일·백엔드·어드민·인프라)

2026.02 출시 · iOS/Android 운영 중

- AI 코치·운동/식단/인바디 트래킹·커뮤니티를 담은 앱. React Native 레거시를 **Flutter+FastAPI로 전면 재구축**, 온프레미스 →GCP 무중단 이전
- **AI 코치**: LLM 프로바이더 추상화(클라우드↔로컬 풀백) · MCP 툴콜링(운동 생성·요약·검색) · 벡터 RAG · 대화 요약→사용자 장기 메모리
- 음식 사진 비전 분석 · 인바디 결과지 파싱 · 소셜 로그인 3종 · 인앱결제 서버 검증(App Store Server API) · 5개 국어
- **Terraform IaC**로 GCP 인프라 코드화(Caddy 자동 HTTPS-Docker Compose) · 일일 백업 자동화 · React 어드민 콘솔

스택 Flutter · FastAPI · PostgreSQL · OpenAI · RAG · MCP · GCP · Terraform 링크 elanfit.app · App Store / Google Play "ELANFIT"

대표 프로젝트 ③ Conoti — AI 공지 자동화 앱 **LAUNCHED**



이엘디 · 1인 개발 (기획·앱·백엔드·AI·인프라)

2026.06 App Store 출시

- 모임 운영자(강사·학원·동호회)용 — 운영자가 **AI와 대화하면 맞춤 공지를 자동 작성·등록·예약 발송**
- **안전한 AI 실행 설계**: 멀티턴 대화 → 구조화 출력 → 단일사용 확인 토큰 게이트 → MCP 등록 (AI 오발송 구조적 차단)
- **커스텀 MCP 서버 이중 경로**: 내부 서비스 토큰 + 외부 발급 인증키로 ChatGPT·Claude 등 외부 AI도 권한·감사 하에 동일 도 구 호출
- 예약/반복 공지 3중 방어(파일락+상태 CAS+먹등키)로 중복 발송 0 · FCM 푸시 · 4티어 구독(RevenueCat) · 백엔드 pytest 140 통과

스택 Flutter · FastAPI · PostgreSQL · OpenAI · MCP · APScheduler · FCM · RevenueCat 링크 conoti.elanfit.app · App Store "Conoti"

대표 프로젝트 ④ 사내 AI 교육 세미나 — 학습 플랫폼 직접 개발 · 강의 EDUCATION



커리큘럼 설계 · 플랫폼 개발 · 강의 단독

2026 · 4부작 시리즈

- 비개발 임직원 대상: 신경망 기초 → CNN/RNN/트랜스포머 → 생성형 AI → LLM 에이전트·MCP
- PPT 대신 **React+TensorFlow.js 인터랙티브 웹 슬라이드** 직접 개발 — 청중이 브라우저에서 MNIST 신경망을 직접 학습시키고 손글씨 인식 체험
- FastAPI로 LLM 추론 SSE 스트리밍·음성합성(TTS) 라이브 데모 연동, 13종+ 시각화 · 무중단 배포 파이프라인

사이드 프로젝트 SIDE

- **whatif-invest.com** — 주식·코인 적립식 투자 시뮬레이션 (2025.09~운영). AWS 서버리스(Lambda·S3·CloudFront)로 월 \$5 이하 운영, 10개 언어·8개 통화, S&P500 MAP 시각화
- **portfolio.eldsoft.com** — 본 포트폴리오 사이트. Astro 정적 빌드 + AWS S3/CloudFront IaC 구성, AI 생성 아트웍·스크롤 인터랙션 직접 구현

이전 경력 하이라이트 EARLIER

- **KT IPaSDN 트래픽 예측 AI** (다임즈, 2022~23) — Airflow 수집 자동화 + LSTM·GRU·Prophet 3개 시계열 모델 비교 선정, KT 대전인재개발원 대상 사내 AI/ML 세미나 진행
- **보령시 신재생에너지 플랫폼** (2023) — FastAPI 수집 서버 · Airflow ETL · Kafka+Logstash 스트리밍 · Redis 캐싱 · 발전량 예측(과업 외 자발 개발) · 다채널 알람(SMTP/FCM/SMS) · Docker MSA
- **천안시 신재생에너지 플랫폼** (2023~24) — LoRa IoT 수집 · 공공데이터 통합, 경험 재사용으로 개발 기간 30% 단축
- **KT 장비 자동화·Dr.IP / KTOA 설계 주도 / DR 플랫폼(한국전력정보) / 유어라운드·실시간 접수앱(이엘디 외주)**

Contact CONTACT

이메일	aofhd0003@gmail.com
포트폴리오	https://portfolio.eldsoft.com — 프로젝트별 상세 페이지·개선 연혁 수록
서비스	elanfit.app · conoti.elanfit.app · whatif-invest.com
GitHub	github.com/YeoJeongHun1