이윤선

AI Inference/Serving Engineer

yoon7829@gmail.com | ckc5800.github.io | github.com/ckc5800 | taepseon.tistory.com

PROFESSIONAL SUMMARY

연구에서 시작해 인프라를 거쳐 모델 추론 서빙으로 넘어온 6년차 AI 엔지니어입니다.

석사와 경력 초반 연구원 시절에 논문 7편과 특허 2건을 쓰며 모델을 연구했고, 인피닉에서는 3인 인프라 팀으로 사내 첫 Kubernetes 플랫폼을 구축하며 관측성 계층을 맡았습니다.

지금은 AI 모델 추론 서빙을 맡아 기존 TTS 스트리밍 API의 첫 응답을 1.9초에서 1.0초로 줄였고, GPU/CPU 듀얼 엔진 서빙 플랫폼을 단독 설계·구축해 엔진 처리량을 4.5배(0.68→3.09rps) 끌어올렸습니다.

최근에는 RAG와 Multi-Agent 시스템을 만들며 LLM 시스템 엔지니어링으로 영역을 넓히는 중입니다.

MiCo AI (구 에이아이세스) | Manager | Apr 2025 – Present

Qwen3-TTS 플랫폼 구축 (2026.03 ~ 현재) | PM 및 풀스택 단독 개발 (Backend / Frontend / Infrastructure)

▸ 배경: 상담센터를 타깃으로 제안할 자사 TTS 솔루션을 신규 구축 — 기존 시스템 교체가 아니라 제품을 처음부터 만드는 과제. 모델 선정(1.7B→0.6B)부터 아키텍처 설계·최적화·배포·운영까지 전 구간 단독 수행, 고객사 PoC·시연 진행

▸ 시스템 아키텍처: vLLM 추론 엔진 위에 FastAPI 게이트웨이, Redis 캐싱·분산락, React 관리자 대시보드까지 전 구간 설계·개발. 긴 텍스트를 문장 단위로 나눠 첫 문장부터 WebSocket으로 내보내는 스트리밍 구조로 첫 소리까지의 대기를 단축

▸ 기능 확장: 참조 음성 5초로 화자 목소리를 복제하는 기능 구현. OpenAI TTS API 호환 인터페이스로 외부 서비스가 코드 수정 없이 연동

▸ 운영 고도화: vLLM(GPU)·Supertonic(CPU) 듀얼 엔진 — 장애 시 Circuit Breaker가 CPU 엔진으로 자동 폴백. 무중단 롤링 재배포 체계와 온보딩 가이드·운영 런북·장애복구 절차·성능 명세를 갖춰 운영 이관 가능 수준으로 정비

▸ 성과 (KPI) — Throughput: 0.68 → 1.41 rps (FP8·executor 최적화) → 3.09 rps (vLLM 0.24 마이그레이션, +80%). 이후 고도화로 TTFB p50 70ms, 8.6 rps (2026.07 실측)

✓ 모델 경량화(1.7B→0.6B) 후 엔진 재튜닝: burst p95 TTFB 0.5초(동시 15)·1초(동시 30), 동시 스트림 수용 한계 36 실측

✓ 32개 언어 단일 플랫폼 (vLLM 10 + Supertonic 22) | ITN 정규화 정확도 50.1% → 70.9% (검증셋 기준)

▸ 핵심 엔지니어링 7건 — 전면 백색소음 장애(SSE JSON이 PCM으로 재생), CPU 스파이크(ReDoS 가드가 치환마다 재생성), 팝 노이즈(PCM 홀수 바이트 정렬), 스토리지 누수, 세션 보안, ITN 토크나이저, WAV 헤더 단일화 — 원인 규명 과정은 포트폴리오

▸ 오픈소스 기여 (vLLM-Omni) — abort된 요청의 sender 상태 미회수를 호스트 메모리 선형 증가의 원인으로 규명. 설정 4종에서 ~45 KB/abort로 일정함을 근거로 인과를 좁히고, 검증 패치로 257 → 60 MiB/h (77% 감소)

▸ 기술: Python 3.12, FastAPI, WebSocket, vLLM-Omni, Redis, React 18, TypeScript, CUDA, FP8 Quantization

TTS 스트리밍 API 리팩토링 (2025.09 ~ 2025.10)

• 기존 TTS(ZipVoice 기반)의 TTFB가 텍스트 길이에 비례해 늘어나 실시간 상담에 부적합했고, 기존 담당자에

게 인수인계 받아 스트리밍 API 리팩토링을 담당 (기여도 100%)

• 브라우저 decodeAudioData()가 두 번째 청크부터 실패하는 고객 이슈를 추적 — 엔진이 첫 청크에만 RIFF WAV 헤더를 포함하는 것이 원인. 모든 SSE 청크에 독립 WAV 헤더를 부착해 해결

• 전체 합성 완료 후 분할 전송하던 기존 구조의 TTFB 한계를 확인하고, gRPC server streaming 기반 실시간 엔드포인트를 신설 — 문장 단위 합성 즉시 전송. API v2 표준화(v1 하위 호환)·Swagger 문서 정비·비교 데모 페이지 제공

• 성과: 같은 텍스트 완주 비교에서 TTFB 1,943ms → 985ms (약 2배 단축, 데모 페이지 콘솔 로그) | Python, FastAPI, Streaming API, ZipVoice

STT 배포 (2025.07 ~ 2025.09)

• AIG 상담센터의 대량 상담 전화를 실시간 처리해야 했으나 동시 채널이 부족 — 팀 내 배포·서빙 담당으로 Faster Whisper 기반 STT 엔진을 맡아 파일 배치와 실시간 스트리밍을 모두 지원하도록 정비

• Triton Python 백엔드로 파일 기반(화자분리 결합)·실시간 스트리밍 두 추론 모델을 구성하고 HTTP·gRPC 제공. 화자 구간을 추론 파라미터로 넘겨 화자별 전사를 생성. 계층 모듈화·컨테이너 자동 재시작으로 안정성 보강 — 고객사 AIG 요구사항 4건 해결

Speaker Diarization (2025.04 ~ 2025.07)

• 상담 품질 분석을 위해 상담사·고객 발화를 구분해야 했음 — Pyannote 사전 학습 모델을 공개 화자분할 데이터로 파인튜닝해 DER로 검증하고, STT와 연동해 화자가 구분된 전사 파이프라인을 구성 (Pyannote, PyTorch, Audio Processing)

인피닉 (INFINIIC) | AI Engineer · MLOps Engineer | Aug 2022 – Apr 2025

Kubernetes 기반 AI 인프라 (2024.03 ~ 2025.04)

• AI 서비스 20여 개를 수동 배포해 배포 1회에 수 시간이 걸리던 상황 — 인프라 엔지니어 3인이 온프레미스에 사내 최초 Kubernetes 플랫폼을 구축하고, 그중 관측성 전 계층(지표·경보·로그)과 스토리지를 맡아 설계·구축·운영

• 스토리지 담당 — NFS 기반 PV/PVC로 모델 가중치·데이터 영속 볼륨을 구성해, 파드가 재시작돼도 모델을 다시 받지 않도록 정비. 클러스터 네트워크(MetalLB LoadBalancer, Nginx Ingress 라우팅)는 팀에서 구성

• 서비스를 Docker로 컨테이너화해 Deployment·StatefulSet으로 올리고, GitLab → Jenkins·Argo Workflow → Harbor → ArgoCD GitOps 파이프라인으로 20개 이상 ML 모델을 자동 배포 (팀 성과)

• 관측 스택 5종을 단독 구축 — Prometheus 지표와 Grafana 대시보드로 20여 개 모델 상태를 한 화면에 모으고, AlertManager 경보로 사람이 발견하기 전에 장애를 알리도록 전환. 온프레미스라 Elasticsearch·Loki를 클러스터에 직접 설치·운영

NLP-based Text Summarization System (2024.01 ~ 2024.07)

• 대화 요약은 BART, 문서 요약은 T5로 나눠 개발. 한국어 대화-요약 쌍 수집·전처리부터 임베딩 모델 선정까지 학습 파이프라인을 직접 구축하고, 사전 학습 표현을 훼손하지 않는 R3F 정규화로 파인튜닝

• ROUGE와 사내 인원 휴먼 평가를 품질 지표로 사용. 사내 메신저에 실제 적용해 긴 공지를 한 줄로 줄이고 업무 대화를 2초대에 요약 (BART, T5, Hugging Face, PyTorch)

Research — Generative Model · 3D Segmentation (2022.08 ~ 2023.12)

• 국방 도메인은 보안 제약으로 학습 데이터 확보가 어려워, Latent Diffusion으로 합성 데이터를 생성하고 실제 모델 학습에 투입해 데이터 부족을 보완할 수 있음을 검증 — 제1저자 논문 2편, 국방기술학회 우수논문상 (2023.11)

• 자율주행 3D 인식을 위해 LiDAR·카메라 센서 퓨전 기반 Trans-Unet을 연구 — 한국자동차공학회 제1저자 논문 2편, 제1발명자 특허 2건 등록 (Latent Diffusion, GAN, Trans-Unet, PyTorch)

이든티앤에스 (EDEN T&S) | AI Engineer | Apr 2022 – Aug 2022

• OCR 데이터 자동 생성 툴 — 수동 주석 병목을, 이미지 로딩·라벨링·저장을 갖춘 어노테이션 툴을 단독 개발해 자동화. 인식 성능 개선으로 프로젝트 성공 성과금 수령 (PyQt, Tesseract, ViT)

큐헷지 (QUHEDGE) | Intern | Sep 2020 – Sep 2021

• 금융 데이터 크롤링·정제·검증·저장까지 자동화 파이프라인을 구축해 수작업 수집을 대체 (Python, Pandas, SQL)

KISTI | 알바 | Oct 2017 – Jan 2018

• 한국어 고어(古語) 데이터 정제 및 전처리 작업 수행

PERSONAL PROJECTS

LLM 시스템 4종 | Jul 2026 – 진행 중 | github.com/ckc5800

• **rag-agent** — LangGraph Corrective-RAG + 하이브리드 검색(FAISS·BM25 RRF). 51문항 유형별 평가셋으로 검색/생성 분리 측정, 채점기 결함을 스스로 검출해 6%p 보정

• **portfolio-mcp** — 포트폴리오 조회용 MCP 서버·클라이언트 양쪽 구현 (FastMCP, 의존성 2개)

• **pdm-agent** — 센서 이상탐지 + LLM 진단 파이프라인. NASA C-MAPSS 검증 — 고장 전 경보 94/100대, RUL 오차(MAE) 12.9사이클(30사이클 전 기준)

• **llm-bench** — 추론 벤치마크 CLI. Ollama 직렬화 병목 진단·재측정 — TTFT p50 12.4초→0.93초

• **공통** — 저장소 8곳에 GitHub Actions CI와 pytest 구성(테스트 함수 300개), 측정 결과는 회귀 테스트로 고정

TECHNICAL SKILLS

Languages: Python (Expert), TypeScript, Bash, SQL | AI/ML · Serving: PyTorch, Hugging Face, vLLM, Triton Inference Server, ONNX Runtime, Whisper

Backend: FastAPI, gRPC, WebSocket, Asyncio | Data: Redis, PostgreSQL, MongoDB, S3, NFS | DevOps: Kubernetes, Docker, GitLab CI, Jenkins, ArgoCD, Harbor, Prometheus, Grafana, Loki, ELK

EDUCATION & CREDENTIALS

M.S. Computer Science, Inha University (2021.08) | B.S. Computer Science, Hannam University (2018.02) | Linux Master, Data Analytics Semi-Professional | English: OPIc IM1

PUBLICATIONS & PATENTS

• 제1저자 논문 7편 — 국방기술학회 우수논문상(2023), 한국자동차공학회 2편, 한국항공우주학회(2023), 스마트미디어학회(2021), ACM(2018), 한국정보기술학회(2018)

• 제1발명자 특허 2건 — 센서 퓨전 기반 시맨틱 세그멘테이션·3차원 해석 방법 (등록 10-2538225 / 10-2538231)