김민석

백엔드 엔지니어

xornjsrlaals@gmail.com 대한민국 GitHub LinkedIn

복구 가능한 실시간 스트리밍과 대규모 비동기 처리 파이프라인을 개발해 온 백엔드 엔지니어입니다. Redis, gRPC, Kubernetes 기반 시스템의 운영 가시성과 처리 성능을 개선했습니다.

Work Experience

네이버클라우드 Applied AI Service

2025.07–2026.01

체험형 인턴

LLM 기반 헬스케어 및 교육 서비스용 채팅 에이전트를 개발했습니다.

Python · FastAPI · MySQL · Redis · gRPC · Kubernetes · MLflow · LangGraph

- Redis Streams와 gRPC-SSE 스트림을 디커플링해 자동 복구 메커니즘을 구축하고, 에이전트 채팅 관련 QA 이슈 0건과 평균 대화 턴 수 2배 증가를 달성했습니다.
- NCP MLflow 커스텀 플러그인과 LangGraph 추적 환경을 구축해 에이전트 운영 가시성을 확보하고 장애 대응 시간을 60% 단축했습니다.
- ContextVar 기반 동시성 제어 구조를 개선해 동시 요청 처리 성능을 최적화하고 인프라 운영 비용을 절감했습니다.

리턴제로 Speech Solution

2024.09–2024.12

학부 ICT 인턴

사내 대량 메일 서비스와 STT OpenAPI SDK를 개발했습니다.

Redis · MySQL · Python · FastAPI · Kubernetes · AWS SES · RQ · Go · gRPC

- Redis와 RQ 기반 비동기 메일 발송 파이프라인에 5단계 Lifecycle 관리와 DB 트랜잭션을 적용해 운영 가시성과 데이터 무결성을 확보했습니다.
- Redis Sorted Sets와 Lua Script 기반 슬라이딩 윈도우 Rate Limiter를 구현해 1만 건 발송 시간을 10분에서 3분 20초로 단축하고 AWS SES 쓰로틀링 발생률 0%를 달성했습니다.
- Transport 패턴으로 JWT 인증을 추상화하고 병렬 파일 스트리밍 파이프라인을 최적화해 대용량 파일 전송 시간을 약 40% 단축했습니다.

Projects

자연어 질의 영상 구간 검색 서비스

2024.03–2024.07

백엔드 개발

Moment Retrieval 모델을 활용한 실시간 영상 구간 검색 및 조합 서비스를 개발했습니다.

Go · Triton Inference Server · gRPC · FFmpeg · Asynq · Redis · MySQL · AWS S3

- AWS S3와 로컬 디스크를 결합한 pre-fetching 전략으로 네트워크 I/O를 최소화해 추론 서버 대기 시간을 단축했습니다.
- Triton 추론 최적화와 영상 ID 중심 구조로 서버 간 페이로드를 줄여 end-to-end 응답 시간을 42% 감소시켰습니다.

Open Source

RQ 기여자

RQ의 기본 로깅 설정이 사용자 로깅 설정을 덮어쓰는 문제를 해결했습니다.

Pull Request #2169

base-repository Product Owner & Maintainer

Python SQLAlchemy 기반 Repository 패턴 제공 라이브러리를 개발하고 유지보수합니다.

GitHub

Education

경희대학교 컴퓨터공학

2019.03–2025.08

학사 · 3.62/4.3

- SW융합대학 학술 동아리 '다솜' 회장 (2023.01–2023.12)
- 경희대학교 유비쿼터스 컴퓨팅 연구실 학부연구생 (2022.11–2023.03)

Training

소프트웨어 마에스트로 14기

2023.04–2023.11

LLM 기반 영상 레시피 서비스 '레시피 에이드'를 개발했습니다.

Skills

Languages	Go, Python
Backend & Data	FastAPI, gRPC, Redis, MySQL, RQ, Asynq
Infrastructure & AI	Kubernetes, AWS S3, AWS SES, MLflow, LangGraph, Triton Inference Server