



# Gijin Kim (김기진)

abjin@naver.com | github.com/abjin | velog.io/@abjin

## EDUCATION

### Dankook University (단국대학교)

Mar 2020 – Present

#### B.S. in Mobile Systems Engineering (모바일시스템공학과), 재학중

- Total GPA: 4.07 / 4.5 | Major GPA: 4.13 / 4.5
- Credits Taken: 126/140 (Expected graduation: Aug 2026)
- Relevant Coursework: 자료구조, 데이터베이스, 확률 및 통계, 컴퓨터 구조, 시스템 프로그래밍, 패턴인식, 컴퓨터 비전

## HONORS & AWARDS

- 성적우수장학금 (학년 수석) | 2026년 1학기
- 단우장학금 | 2025년 2학기

## WORK EXPERIENCE

### Nudge Healthcare (Cashwalk)

Apr 2022 – Apr 2024

#### Back-end Part Leader & Developer

스크럼 진행, 코드 리뷰, API 설계, RDB/NoSQL 최적화 등 프로젝트 전반의 기술적 리딩과 팀 관리를 수행했습니다.

#### 실시간 대규모 채팅 서버 개발 (캐시워크, 모두의 챌린지)

- 일간 활성 사용자(DAU) 40만 규모의 대용량 트래픽을 처리하는 채팅 서버 아키텍처 설계 및 개발
- Artillery 기반 부하 테스트를 통해 대규모 동시 접속 환경에서 메시지 처리 안정성 검증
- 읽음 처리, 1대1 채팅, 단체 채팅, 채널 채팅(500만명 이상 동시 발송) 등 다양한 타입의 메시징 기능 개발
- Node.js, NestJS 기반 API 설계 및 Socket.io를 활용한 실시간 통신 구현
- 멀티테넌트 아키텍처를 적용하여 시스템 확장성 확보
- RDB/NoSQL(MySQL, Dynamo DB, Redis) 설계 및 최적화를 통해 데이터 처리 성능 개선

#### 글로벌 캐시워크 백엔드 개발 (EU/US/UK)

- 북미/유럽 대상 글로벌 서비스 백엔드 시스템 설계 및 개발
- 글로벌 환경에서 네트워크 지연 및 데이터 일관성을 고려한 안정적인 아키텍처 구축
- AWS(EC2, RDS, S3, SQS, Lambda 등) 클라우드 환경에서 기능 유지보수 및 운영

## PROJECTS

### RAG Pipeline Optimization Experiment

TopK, Chunk Size, Overlap 파라미터가 응답 품질과 시스템 비용에 미치는 영향을 정량적으로 분석하고, trade-off 관점에서 최적 값을 도출하는 실험을 수행했습니다.

- 실험 설계: 3개 파라미터(TopK 1/3/5, Chunk Size 256/512/1024, Overlap 0%/15%/30%)의 27개 조합에 대해 21개 질문 셋으로 3회 반복 실험 수행
- LLM-as-a-Judge 기반 평가: RAGAS 프레임워크(Faithfulness, Answer Relevancy, Context Recall/Precision)와 시스템 지표(Latency, Token Usage) 동시 측정
- Sweet Spot 도출: 실험 결과 분석을 통해 품질과 비용 간 최적 균형점(TopK=3, Chunk Size=512, Overlap=15%) 도출
- 성과: Slack RAG 봇의 Answer Relevancy 약 17% 향상, Total Token 사용량 약 57% 절감

### AI-Powered Slack RAG Bot

팀 내부 지식(GitHub, Notion)에 대해 정확하고 근거 있는 답변을 제공하는 RAG 기반의 Slack 봇을 개발했습니다.

- 질문 분류 및 라우팅: LangGraph를 활용하여 사용자 질문의 의도를 분석하고, 각 목적에 맞는 서비스로 동적으로 라우팅하는 워크플로우를 구현
- Vector Embedding & Search: Gemini Embedding Model을 사용해 GitHub와 Notion의 텍스트 데이터를 벡터로 변환하고, Pinecone 벡터 DB에 저장하여 유사도 검색 수행
- 자동 임베딩 업데이트: NestJS Schedule 모듈을 이용해 주기적으로 GitHub/Notion의 최신 변경 사항을 자동으로 감지하고 Pinecone DB의 임베딩을 업데이트하는 Cron 작업 구현
- Multi-Tenant Architecture: 여러 Slack 워크스페이스가 독립적인 정보를 사용하도록 멀티테넌시를 지원하는 DB 스키마와 로직 설계

**Tech Stack:** NestJS(TypeScript), LangChain/LangGraph, Gemini API, Pinecone, Prisma, MySQL

## AI/ML Study Blog (Velog + GitHub Repo)

Transformer/Attention, CNN, ViT·U-Net, 패턴인식 이론 등 기초부터 주요 주제를 체계적으로 학습·정리한 블로그입니다.

- Transformer & Attention: Seq2Seq 한계, Attention 메커니즘, Transformer 개념 정리
- Computer Vision: CNN 기본, LeNet-5, AlexNet, VGG, ResNet, Batch Normalization, MobileNet, EfficientNet, 다양한 비전 작업, U-Net, ViT
- Pattern Recognition: Regression, Least Squares, MLE, LS vs MLE, Logistic Regression, Gradient Descent & Optimization, Regularization, Model Evaluation, SVM

## Calorie Analysis AI Server from Food Images

YOLO 모델을 파인튜닝하여 음식 객체를 분류하고, 음식 사진 기반으로 영양 정보를 제공하는 AI API 서버를 개발했습니다.

- YOLOv11 모델 파인튜닝: 90종의 음식 클래스 데이터셋으로 모델을 직접 파인튜닝하여, 음식 탐지 정확도를 mAP 0.85까지 향상
- AI 모델 서빙 최적화: 파인튜닝된 YOLOv11 모델을 FastAPI 서버에 통합하여 실시간 이미지 처리가 가능한 효율적인 AI 서빙 파이프라인 구축
- 클라우드 기반 아키텍처: AWS(EC2, S3, RDS)를 활용한 안정적이고 확장 가능한 서버 인프라 설계 및 Docker로 배포 자동화
- API 설계: 이미지 업로드부터 영양 정보 분석 결과 반환까지의 과정을 처리하는 RESTful API 설계

**Tech Stack:** Python, FastAPI, PyTorch, YOLOv11, Docker, AWS (EC2, S3, RDS)

## AI-based Clothing Recycling Platform

YOLO 모델을 서빙하여 의류 재활용 선별 과정을 자동화하는 서비스를 개발했습니다.

- AI 모델 서빙 API 개발: FastAPI를 사용하여 의류 이미지 URL을 입력받아 분류 및 결함 검출 결과를 반환하는 추론 API 엔드포인트 구현
- 모델 최적화 및 서빙: PyTorch 모델을 TorchScript로 변환하여 추론 성능을 최적화하고, 실시간 이미지 처리가 가능한 서빙 파이프라인 구축
- 클라우드 기반 배포 자동화: Docker 이미지를 빌드하여 Docker Hub에 푸시하고, AWS Elastic Beanstalk에 자동으로 배포되는 CI/CD 파이프라인 설계

**Tech Stack:** FastAPI, PyTorch, TorchScript, OpenCV, Docker, AWS Elastic Beanstalk

## RESEARCH INTERESTS

### Retrieval-Augmented Generation (RAG) & Domain-Specific LLM

RAG 기반 봇 구현 경험을 바탕으로, 외부 지식 베이스를 활용해 LLM의 신뢰도를 높이고 환각 현상을 줄이는 연구에 관심을 두고 있으며, 특히 도메인 특화 데이터에 대한 효율적인 검색, 인덱싱, 생성 전략을 탐구하고 있습니다.

### LLM Serving & Inference Optimization

LLM 추론 과정에서 발생하는 병목과 처리량 한계를 주요 문제로 인식하고, 추론 파이프라인과 서빙 구조에서의 최적화 문제를 탐구하고 있습니다.