PORTFOLIO

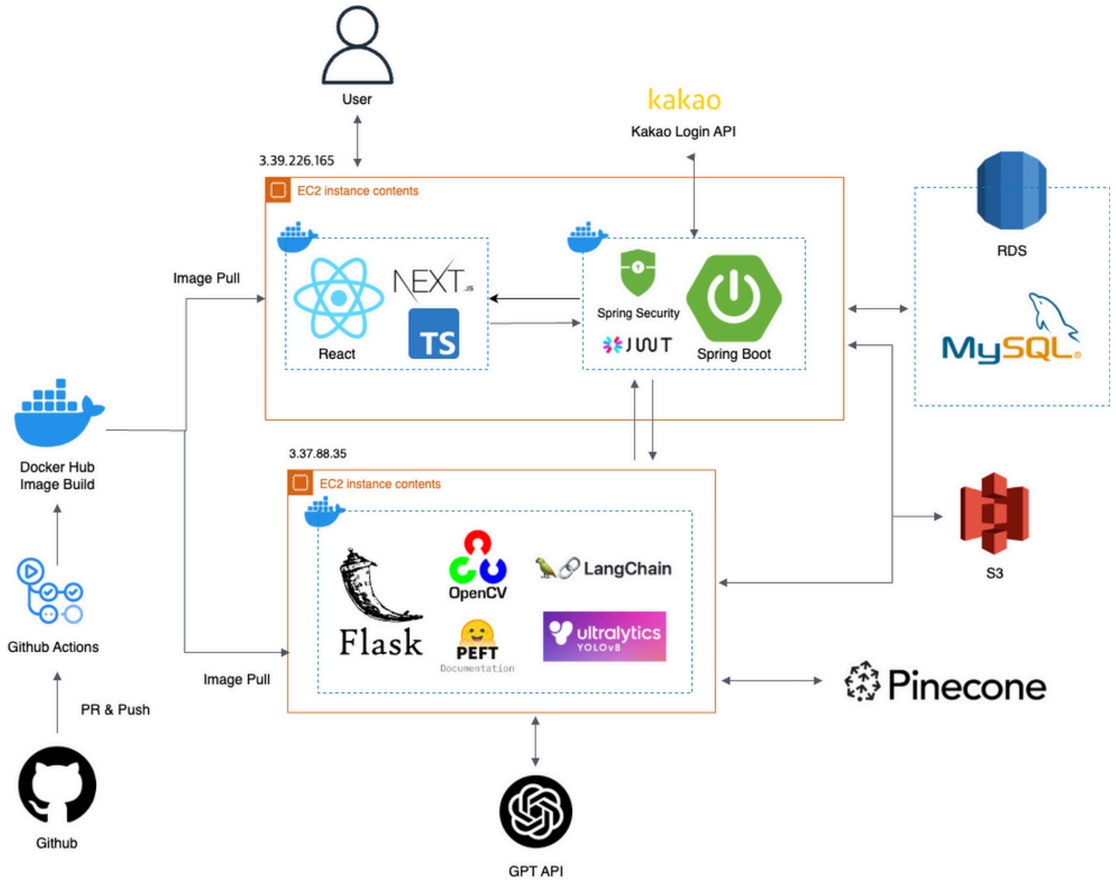
M E & P R O J E C T S

Think flexiblely, Design deeply, Solve precisely

AI/ML Engineer MINKYUNG KIM

2025

StarBridge | LLM & Generative Vision AI-based Pet Loss Support Service



Overview

- **Led the development of an AI-based service to support users experiencing pet loss, integrating LLMs, vision AI, and emotionally engaging UX.**
  - 🔍 **1. Chatbot for Expert Guidance (GPT, RAG)**
    - Built a chatbot offering tailored advice (e.g., senior pet care, insurance, funerals)
    - Filtered irrelevant queries by dropping low-similarity results via Pinecone vector DB
    - Increased precision by chunking policy documents, extracting keywords (GPT-API), and applying similarity-based retrieval
  - 🌌 **2. Pet Constellation Generator (YOLO, SAM, PidiNet)**
    - Created constellations from pet images, helping users preserve memories in a symbolic form
    - Addressed segmentation issues (e.g., mis-splitting, imprecise areas) using YOLO + box-based SAM for cleaner UI/UX
  - 💌 **3. Personalized Letters via Generative AI (Dreambooth - Stable Diffusion)**
    - Delivered postcards with pet images and letters “written” by the pet
    - Used stylized LoRA to reduce the uncanny valley and improve emotional quality
  - 🚀 **4. Deployed all the ML Pilpeline on AWS GPU server (Gitub CI/CD, Docker)**
    - enabled ML API on g4dnx.large instance with CUDA settings

StarBridge | LLM & Generative Vision AI-based Pet Loss Support Service

Chatbot for Expert Guidance
(GPT, RAG)
Guide informative contexts of
senior pet care, insurance, funerals

Personalized Letters via Generative AI
(GPT, Dreambooth – Stable Diffusion)
Generate heartfelt letters using GPT
based on users' memory posts about their pets.
Create personalized images
that reflect the pet's appearance.

Pet Constellation Generator
(YOLO, SAM, PidiNet)
Automatically segment the pet from a photo, extract contour lines, sample key points, and connect them using a Minimum Spanning Tree (MST) to form a constellation.

https://github.com/starlight-dari
https://www.youtube.com/watch?v=iklY5DqPYYw

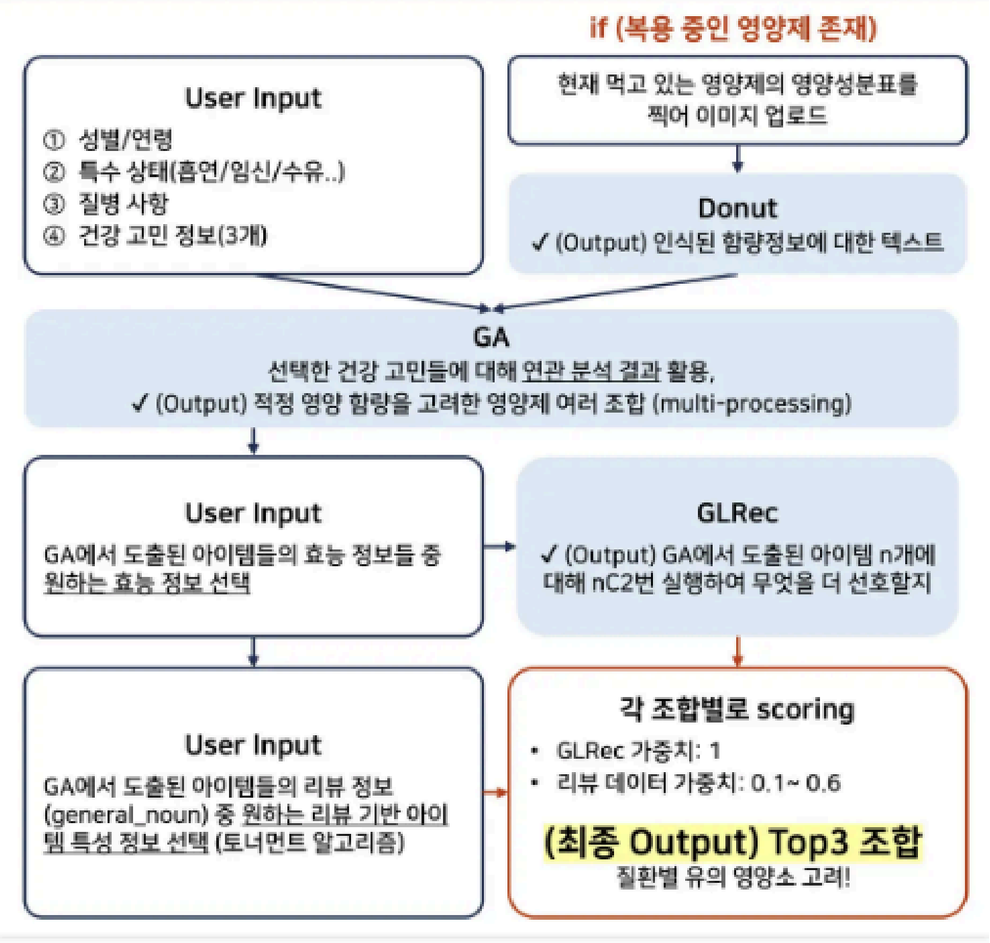
PillForMe | Personalized Supplement Recommendation App (LLM + Genetic Algorithm)

RTX4090 GPU(24G)
32개 CPU의 로컬 환경에서 구현
total running time: 2~3 min

Flask
BE: flask(port:5000)
Donut_MLserver: flask(port:6000)
GA_MLserver: flask(port:7000)
GLRec_MLserver: flask(port:8000)

Pinecone
리뷰 데이터 활용 시 VectorDB 활용

FE: react native expo(port:8081)



Overview

- **Led the development of a personalized supplement recommendation app leveraging LLMs, and graph-based prompt engineering.**
  - 🧬 **1. Custom Recommendation Model Design (LLM + GA)**
    - Identified limitations in existing systems that couldn't recommend multi-product combinations based on individual user health profiles
    - Co-designed a novel model integrating OCR, genetic optimization, and LLM reasoning to tailor complex supplement bundles
  - 🔄 **2. Architecture Pivot & Data Strategy**
    - Initiated a structural shift from review-based similarity to LLM-based recommendation, based on academic research findings
    - Persuaded the team by presenting scenario-driven technical analysis and proposing a new data pipeline
    - Designed a graph-based prompt structure using clustered user-item interaction data for enhanced LLM context
    - Also complemented GA algorithm with multi-threading
  - 🧪 **3. Model Fine-tuning & Performance Improvement**
    - Fine-tuned LLaMA model using LoRA, and resolved overfitting via data complexity control and reprocessing
    - Achieved 98% recommendation accuracy, enabling customization based on gender, age, concerns, and dosage
  - 📱 **4. Implemented a prototype app with React Native**
    - Connected the APIs for each ML Server (OCR/GA/LLM) and implemented FE with React Native

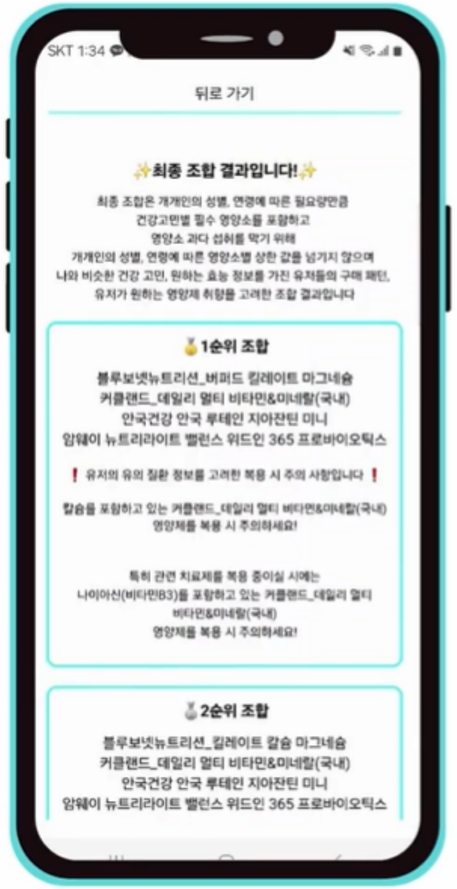
2024

PillForMe | Personalized Supplement Recommendation App (LLM + Genetic Algorithm)

1. Generate candidate supplement combinations using a Genetic Algorithm (GA) based on user inputs (e.g., health concerns, lifestyle, demographics).

2. Select desirable combinations using tournament-based filtering with:

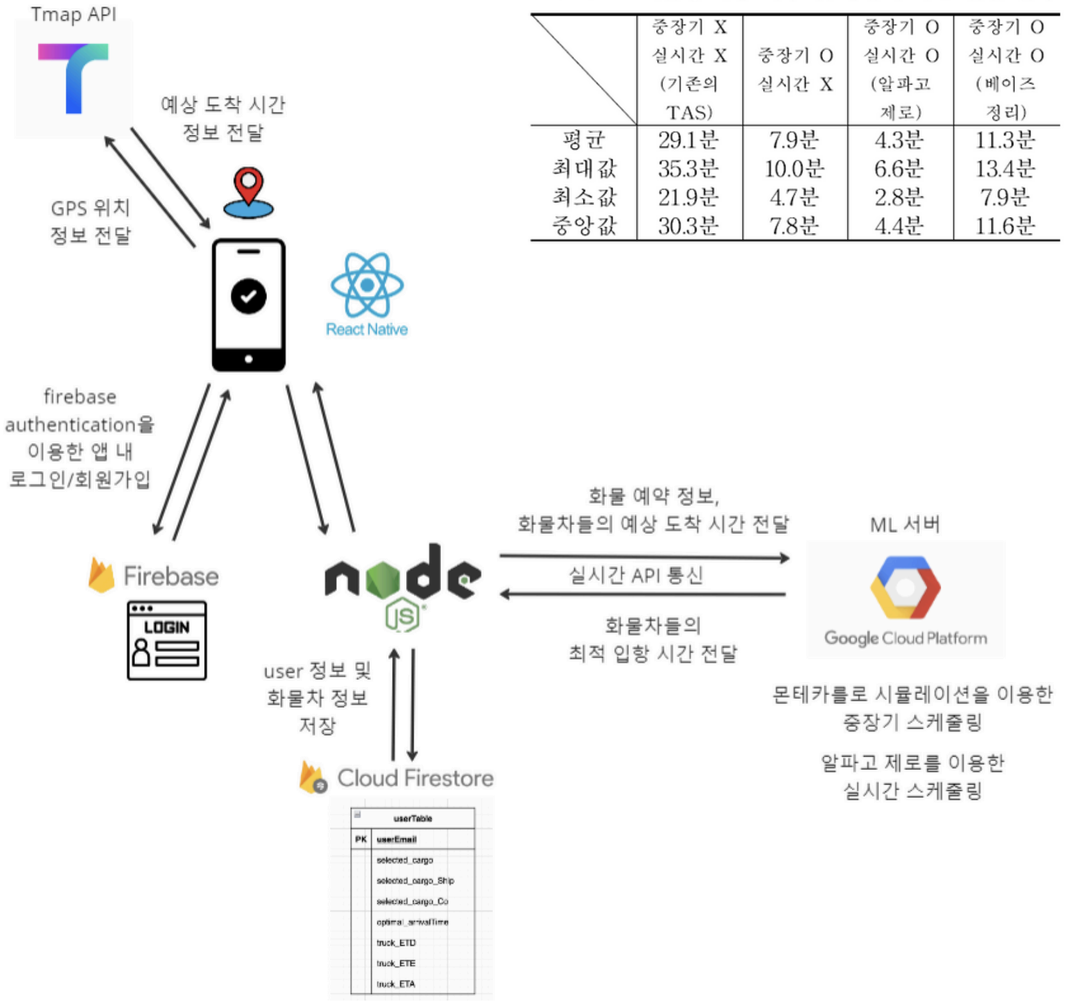
- User-specified functional goals
- Review-based product characteristics
- Pairwise LLM-based recommendations across all possible combinations (nC2)



3. Score all combinations using weighted factors (e.g., GLRec score, label data relevance), **and return the Top 3 optimized supplement bundles.**

https://github.com/PillForMe
https://www.slideshare.net/slideshow/20-boaz-19-boaz-pill-az-pill-for-me/270919517

Port Scheduler | Reinforcement Learning-Based Truck Scheduling System



3. 실시간 스케줄링 모델의 성능 평가

<표 1> 스케줄링 적용 여부 및 종류에 따른 실험 결과

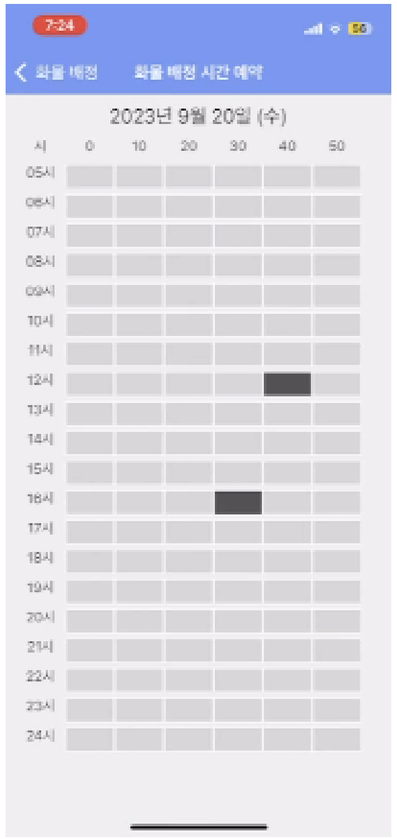
	중장기 X 실시간 X (기존의 TAS)	중장기 O 실시간 X	중장기 O 실시간 O (알파고 제로)	중장기 O 실시간 O (베이즈 정리)
평균	29.1분	7.9분	4.3분	11.3분
최대값	35.3분	10.0분	6.6분	13.4분
최소값	21.9분	4.7분	2.8분	7.9분
중앙값	30.3분	7.8분	4.4분	11.6분

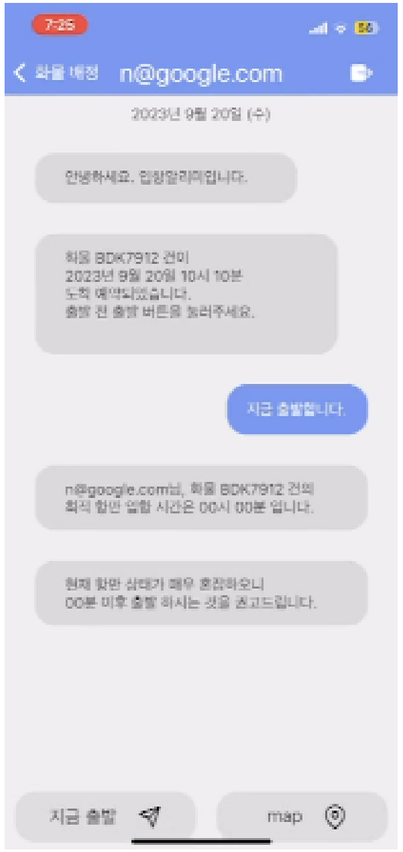
Overview

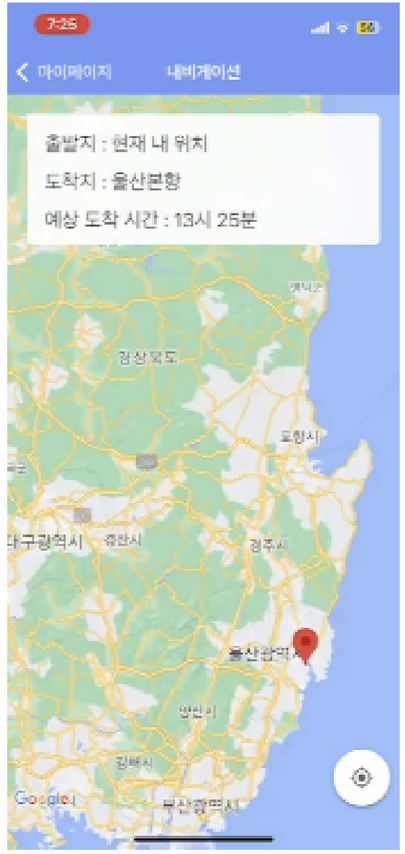
- **Built a port truck scheduling system using Monte Carlo simulation, AlphaGo Zero-style reinforcement learning, and Bayesian inference to reduce congestion at Ulsan Port.**
  - 🧱 **1. Hybrid Scheduling Model (Monte Carlo + RL + Bayes)**
    - Designed long-term and real-time models to optimize truck arrivals
    - Used 2 years of data to estimate reservation capacity via Monte Carlo simulation
    - Applied reinforcement learning with MCTS and Bayesian updates for dynamic real-time rescheduling
  - 🔁 **2. Strategic Pivot & Implementation**
    - Switched from an underperforming graph-based model to an RL-based approach, inspired by Go AI
    - Led team transition by proposing new structure, guiding daily discussions, and formulating RL strategy
    - Implemented final model as a Flask ML server with API integration
  - 📊 **3. Results & Deployment**
    - Reduced average wait time by 85% (29.1 → 4.3 mins), with ~30M KRW annual savings
    - Built full-stack app: Flask (ML), Node.js (backend), React Native (frontend)
    - Conducted on-site visits and managed team workflow to ensure project delivery

2023

Port Scheduler | Reinforcement Learning-Based Truck Scheduling System







1. Time Slot Reservation (Long-Term Scheduling)
Fully booked time slots are closed to new reservations. Truck capacity per slot is limited by an algorithm.

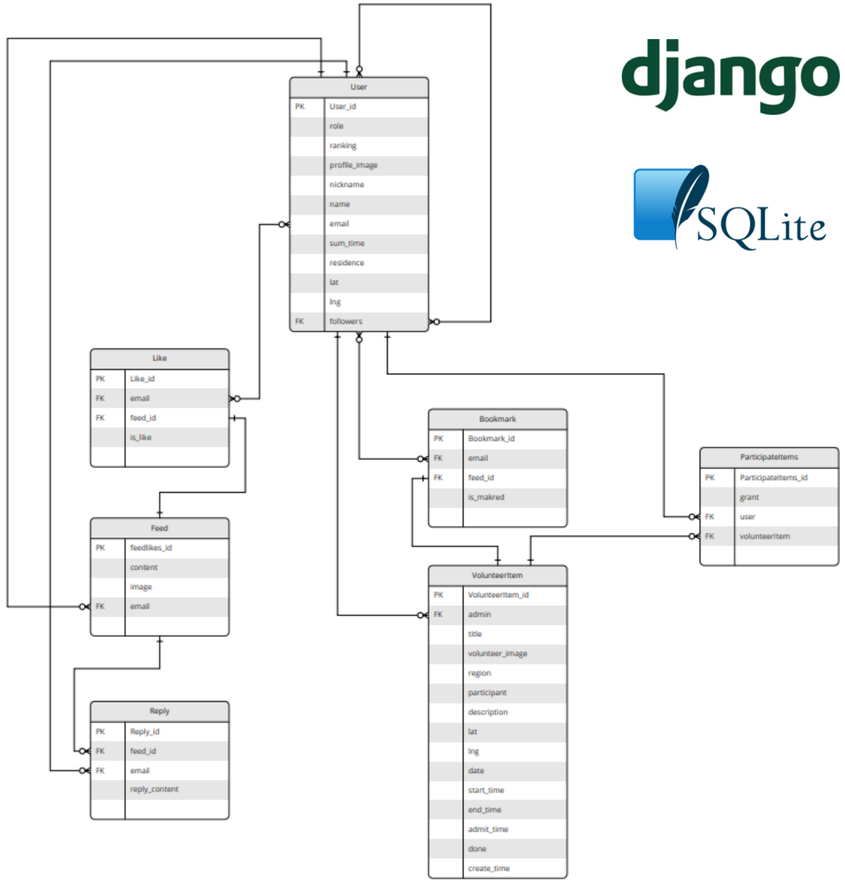
2. AI Delay Guidance (Real-Time Scheduling)
When the driver taps "Depart", AI recommends a delay (0–30 mins) based on current traffic and scheduling conditions.

3. Live GPS & Queue-Aware Scheduling
Google Maps displays real-time location via GPS. AI uses this to track queue status and reschedule arrivals dynamically.

https://www.youtube.com/watch?v=FlmLo69S6d4
https://kiss.kstudy.com/Detail/Ar?key=4059591

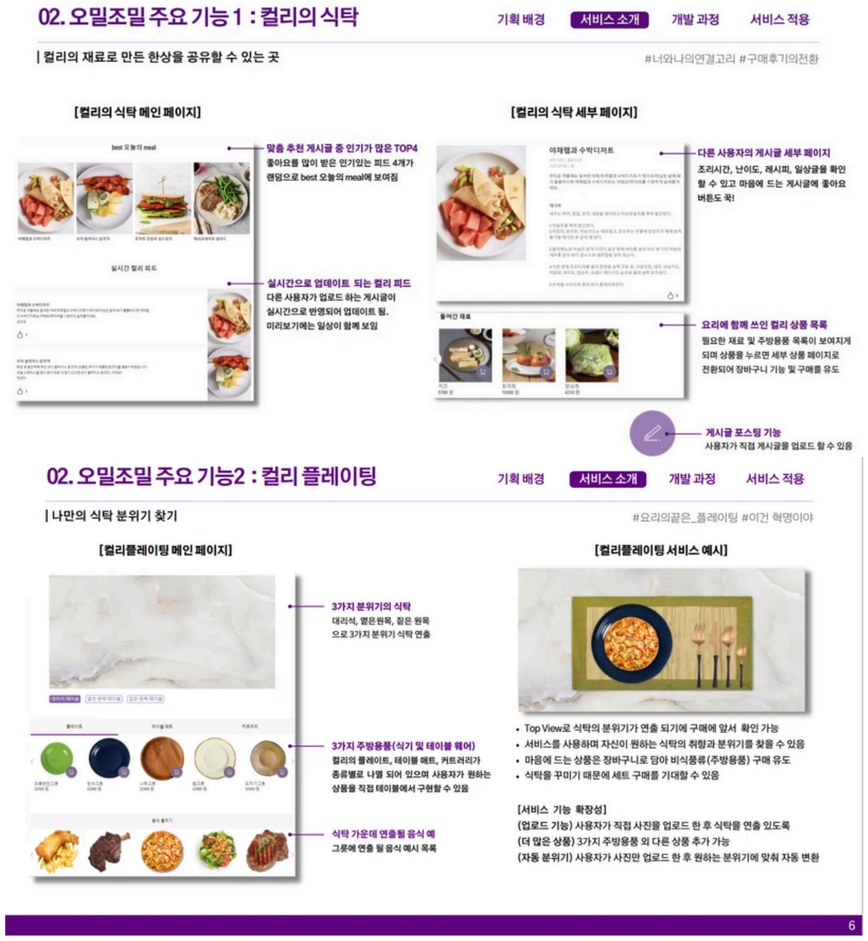
++ Frontend / Backend / DB Experiences

https://lean-advantage-d78.notion.site/45ad43382ff4467db8f1091c10f0c499



Developed core volunteer platform features
including user auth, CRUD, social functions, and distance-based
recommendations using Kakao Map API.
Also structured ERD for SNS-like features
and managed data withSQLite DB

https://lean-advantage-d78.notion.site/s-339970fb06d442dab4be92c3c6ccd8c2



Implemented image upload/view using multipart/form-data and
built key UI pages including login, signup, carousel-based main,
detail views, cart, and post creation.

2022

THANK YOU

THAT'S MY PORTFOLIO SO FAR.