



최인규 포트폴리오 (PDF)



About Me

AI 모델을 실서비스에 통합하고 안정적으로 운영하는 것에 집중하는 백엔드 개발자입니다.
RAG, 온디바이스 AI, 실시간 추론 서버 구축 경험을 보유하고 있습니다.

Stack

AI / ML | Python · PyTorch · llama.cpp · FAISS
Infra | FastAPI · MongoDB · Docker · AWS EC2
Tools | Git · Notion

Contact

Mail. chldlsrb08@naver.com
GitHub. <https://github.com/yyyyeengyu>
Blog. <https://yyyyeengyu.tistory.com>
Phone. 010-8269-9915

팀 프로젝트

프로젝트 개요



DOGTOR

2025학년도 2학기 캡스톤
프로젝트
온디바이스-AI기반 **애완견**
케어 서비스

Date

2025. 08 ~ 2025.12 [4개
월]

Stack

BE - FastAPI
AI - Python · llama.cpp ·
Qwen2.5-7B-Instruct
Infra - EC2 · Docker

Teammates

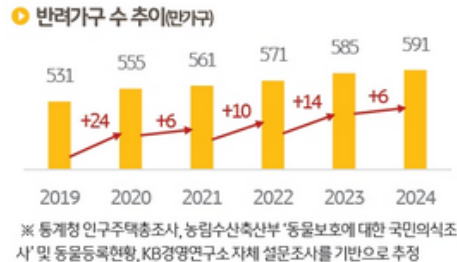
- 5인 팀 개발
- 최인규 - AI · AIBE · Infra
· PM
- Design · Web Frontend
1
- App Frontend 1
- Backend 1
- AI 1

Github Repository

[백엔드 리포지토리](#)
[AI 리포지토리](#)

More

Dogtor는 'OceanLight'사와
협업하였습니다.



현재 동물병원 시스템은 반려견의 가벼운 질병에도 전문기관을 찾아야 하며, 이에 대한 효율성이 떨어집니다.

반려동물을 기르는 반려인들은 **연도에 비례하여 증가하는 추세**를 보이고 있지만, 이는 위의 단점이 더욱 부각됩니다.

초보 반려인들은 반려동물에 대한 관심과 애정이 넘치지만, 이에 비해 **반려동물의 특징과 행동특성을 잘 알지 못하는 경우가 허다**합니다.

이 점에서 **빠르게 애완견을 케어해줌과 동시에 동물병원 연계까지 가능한 Dogtor**를 기획하였습니다.

프로젝트 기여 내용온디바이스-AI 개발

- 산학연계 회사 과제내용 완수를 위하여 온디바이스AI 개발
 - llama-cpp를 활용한 모델 양자화 경험
 - 7B 모델 (약 15GB) → 4bit GGUF 양자화 성공 (4.1GB)

데이터셋 전처리

- 실제 수의사 대화 **QA데이터 전처리 · 증강 진행**
 - 중복된 데이터 삭제 · 질문 paraphrasing 진행
 - **기존 데이터 2000개 → 4배 증강**
 - 랜덤 관련 증강이 모델 학습에 도움이 되지 않았음
 - paraphrasing과 역번역으로 증강 다시 진행, 모델 학습 완화

모델 Lifespan 설정

- 서버 테스트 시 **온디바이스AI와 유사한 지연율의 API 필요**
 - FastAPI 기반 inference 서버 구축
 - 서버 startup 단계에서 **LLM모델 preload**
 - API 요청마다 모델 재로드 방지
 - **Inference latency 감소 및 안정적인 응답 처리**

또한, 2026 AI Expo Korea
국제인공지능대전에
3일간 출품되었습니다.

트러블슈팅증강 데이터로 학습 시 정확도 감소상황

데이터 증분을 위해 랜덤 삽입 · 제거 증강을 시도하였으나, 오히려 모델의 답변 정확도가
떨어지고 문맥이 어색해지는 상황이 발생하였습니다.

```
{
  "question": "우리 개가 먹은 걸 다 게워내요",
  "answer": "• 주요 원인: 과식·급섭취, 감염성 위장염, 이물 또는 독",
},
{
  "question": "게워내요 우리 먹은 걸 다 개가",
  "answer": "• 주요 원인: 과식·급섭취, 감염성 위장염, 이물 또는 독",
},
{
  "question": "개가 우리 걸 먹은 다 게워내요",
  "answer": "• 주요 원인: 과식·급섭취, 감염성 위장염, 이물 또는 독",
},
{
  "question": "걸 먹은 개가 우리 다 게워내요",
  "answer": "• 주요 원인: 과식·급섭취, 감염성 위장염, 이물 또는 독",
},
}
```

특히 랜덤 삽입 · 제거 기법을 적용한 후의 답변이 문맥이 심하게 깨졌고, **사용자가 알아보지 못할 정도의 문장을 구사하는 현상**이 발생하여 대처해야 하였습니다.

해결방향

단순히 데이터를 늘리는 방법보다는 문맥 정확도까지 챙겨야 하는 상황에 **랜덤삽입 · 제거 증강은 적합하지 않다고 판단**,
입력의 문맥과 의미를 살릴 수 있는 증강 방식을 선택해야 했습니다.

랜덤 관련 기법들은 **과감하게 배제하고**, 문장 **paraphrasing**과 **역번역 증강**을 선택하여 다시 학습을 진행하였습니다.

결과

문장의 의미와 **문맥이 깨지는 문제를 해결함과 동시에 데이터 증분까지 성공**하며 모델의 답변이 질문 의도와 더 자연스럽게
연결되는 것을 확인하였고, **사용자 입장에서 납득 가능한 출력들**을 내뱉는 것을 재차 확인하였습니다.

Fresh Money

2025학년도 1학기 캡스톤
프로젝트
청년들을 위한 **청년정책** 및
재정관리
서비스

Date

2025. 03 ~ 2025. 07 [4개
월]

Stack

BE - FastAPI · MongoDB
AI - Python · LangChain ·
OpenAI API
Infra - AWS EC2 · Docker ·
NginX

Teammates

- **최인규** - AI · AIBE · Infra
- Web Frontend 2
- Design 1
- Backend 1
- AI 1

Github Repository

[AI 리포지토리](#)

More

[Fresh Money 블로그](#)

프로젝트 개요

25세 A씨 수입, 지출 및 자산 현황 (단위 : 원)			자료 : 금융감독원
월 수입		250만	
월 지출 (168만)	• 고정비 (0)	-	
	• 변동비 (88만)	식비 30만, 용돈 40만, 통신비 40만, 교통비 12만, 미용비 2만	
	• 저축 (80만)	청약 10만, 청년도약계좌 70만	
연간 비용		600만	
자산 (540만)		입출금 통장 500만, 청약저축 40만	
부채 (1400만)		학자금 대출 잔액 1400만	

프로그래밍 관련 사회초년생들의 초기 연봉은 2,500 ~ 3,000에 웃돌고 있습니다.
사회초년생이 첫 월급을 받을 때, 초기에는 어떻게 관리해야 하는지 고민하는 상황이 자주 발
생합니다.

이 점에서 사회초년생의 재정관리를 쉽게 풀어나가게 해주고자 Fresh Money를 기획하게 되
었습니다.

핵심 기능자산 증식

- **투자 추천** : 간단한 입력으로 사용자에게 **가장 적합한 투자 방식** 추천
- **금융상품 추천** : 사용자의 재정상태에 맞게 **제1금융권의 10여개 금융사의 금융상품** 추천
- **지원정책** : 사용자의 **지역과 재정상태에** 맞는 **청년지원정책** 추천



투자 추천 · 지원정책 추천은 신뢰도 있는 추천을 위해 **RAG기법**을 도입하였습니
다.

자산 유지

- 신용등급 분석 : **공식 신용등급 계산법 (KCB, NICE지키미)**을 통하여 현재 신용등급과
신용등급 관련 도움 제공
- FM 챗봇 : 청년분들이 헛갈려하실 수 있는 **자산 관련 정책** 등등을 **편하게 질문하실 수 있**
도록 챗봇 제공

프로젝트 기여 내용RAG 구축

- 제1금융권에서 제공하는 은행상품들을 크롤링하여 **임베딩 · 벡터화**
 - LangChain의 RecursiveCharacterTextSplitter를 통한 **문서 청크 분할**
 - FAISS 라이브러리를 사용하여 **벡터검색 활성화**
 - 금융상품 추천 기능에서 주 요소로 작용하게 됨

LangChain기반 모델 파이프라인 구축

- LangChain에서 제공하는 모듈로 RAG 코드 구성
 - LCEL 체이닝을 통한 RAG 파이프라인 작성 및 코드 가독성 향상

- RetrievalQA 체인으로 RAG 파이프라인 캡슐화
- StrOutputParser를 통한 객체 문자열 해독

트러블슈팅 / 이슈프롬프트 기법 Few-shot 도입상황

OpenAI API를 사용하는 모든 기능에서 422에러가 발생하는 상황이었습니다.
모델의 출력은 문자열로 반환되는데, 서버 → 프론트 반환 시 JSON 형태로 반환해야 했지만 문자열을 그대로 반환하여 422오류가 나온 상태였습니다.

해결방향

LangChain에서 JSON을 강제시켜주는 메서드가 존재했었지만, 25년도 초중반의 LangChain은 출력 강조 메서드가 제대로 작동하지 않았습니다. 오히려 다른 오류를 불러올 수 있는 상황이었기에, 차선책인 Few-Shot 프롬프팅을 진행하였습니다.

결과

각 기능마다 프롬프트에 예시 JSON 출력을 2개씩 입력한 후 실제 10번 호출 테스트 시 422오류를 완전히 해결하였고
더불어 Few-Shot 프롬프팅을 통하여 응답이 단일화되어 사용자 입장에서도 보기 편한 응답들을 기대할 수 있었습니다.

프로젝트 개요



LightBox

2025 전국 고등학교 동아리 SW 경연대회 본선 프로젝트
YOLO & MediaPipe
FaceMesh 기반
실시간 졸음 감지 서비스

Date

2025. 09 ~ 2025. 11 [3개월]

Stack

BE - FastAPI · MongoDB
AI - Python · YOLOv11 ·
Mediapipe FaceMesh
Infra - AWS EC2

Teammates

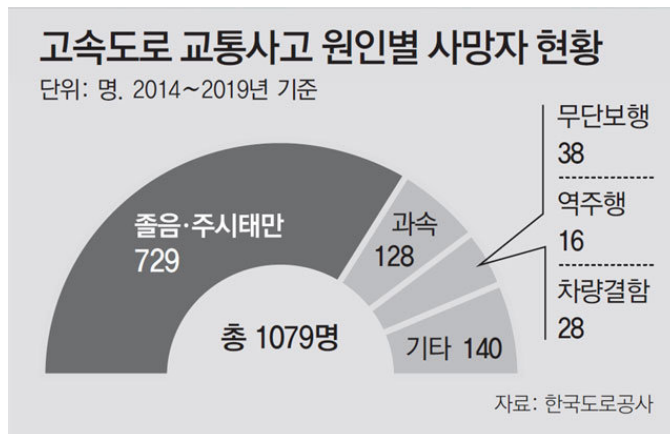
- 최인규 - AI · AIBE · Infra
- Design · Web Frontend 1
- Backend 2
- AI 1

Github Repository

[AI 리포지토리](#)

More

LightBox는 제3회 아진실리경진대회에서
경북교육청 부스에 출품되었습니다.



고속도로 교통사고 원인 중 가장 큰 이유를 차지한 요소는 '졸음운전'입니다. (전체 67.6% 졸음·주시태만으로 인한 사망)

장시간 운전으로 인한 졸음과 주시 태만이 대형 사고를 유발할 수 있다는 점에서 기술적인 대책이 필요하다고 생각하였고,
눈 감김을 감지하고 졸음 유무를 판단하는 서비스를 제공함으로써 교통사고 위험을 줄이고 안전한 도로 환경을 조성하는 데에 기여하고자 합니다.

프로젝트 기여 내용MediaPipe기반 눈 감김 감지

- 졸음운전 감지를 위해 눈 근처 6개 랜드마크 기준으로 감김 퍼센트 계산
 - 유클리디안 계산법 기반 **EAR (Eye Aspect Ratio)** 알고리즘 구현
 - 첫 시작 5초 EAR값을 평균 임계값으로 지정하며 임계값보다 EAR 알고리즘 값이 낮을 시 졸음운전으로 판단
 - 사람마다 다른 눈 폭에도 정상적으로 감지

운전 이후 졸음도 리와인드 제공

- 주행 중 **EAR 기반 졸음 지표 수집**
- 시간 축 기준 졸음도 그래프 생성
- 운전 종료 후 졸음 발생 구간 리와인드 분석 제공
 - 사용자에게 졸음 패턴 피드백 제공

YOLO기반 얼굴·눈 크롭

- 추론 정확도를 높이기 위해 눈 감지 및 crop
 - 사용자 화면에도 보여주며 눈이 감지되고 있는지 확인
 - 내부 테스트 기준, 눈 영역 crop 적용 후 **오탐을 줄이고 실시간 감지 안정성을 개선**

트러블슈팅 / 이슈모델 추론환경 제약 대응상황

학교에서 제공한 EC2에서 AMI 환경이 GPU 드라이버를 호환하지 않아 **CUDA Runtime** 과 충돌한 현상이 발생하였습니다.

이 때문에 EC2의 GPU를 사용하지 못하고 CPU환경에서 추론 환경을 구성해야 하는 상황이 발생하였습니다.

해결방향

기존 EC2 환경에서 여러 방법을 통하여 해결하려는 행동을 보였습니다.
커널 버전 관리, DRM 우회, 드라이버 버전 다운그레이드, NVIDIA Runfile 설치 등 여러 방안을 실행해보았지만
동일한 오류가 반복되어 차선택인 **CPU 기반 추론 구조로 변경**하였습니다.

GPU 가속이 필요한 파트는 YOLO 추론밖에 없었고, mediapipe 자체는 CPU로 구동되어
줄음 감지 자체에는 크게 문제될 게 없다고 판단하였습니다.

따라서 **YOLO의 추론 정확도를 감소하는 대신에, 실시간성을 높이는 방향으로 트레이드오프를 구상**하였습니다.

결과

YOLO 추론 코드 중 imgsz를 기본값보다 낮춰 추론을 진행하여 **CPU 환경에서도 실시간 처리 안정성을 확보**하였습니다.
이로써 **YOLO의 실시간성과 줄음 감지 정확도 두개를 챙겨 정상적으로 작품을 출력**하였습니다.

개인 프로젝트



Swagger Inspector

OpenAPI RAG 및 분석해 주는
챗봇 서비스

Date

2026. 01 ~ 2026. 02 [1개월]

Stack

Agent - Codex
BE - FastAPI · MongoDB
AI - Python · text-embedding-3-small · ChatGPT Codex

Github Repository

[전체 리포지토리](#)

More

프로젝트 개요

스케일이 큰 프로젝트를 진행하다보면 기능들이 매우 많아집니다.

이때 백엔드 코더분들을 제외한 다른 스택의 팀원은 기술명세가 제대로 되어있지 않은 한 어떤 기능들이 구현되어있는지, 요청 방식을 고민해야 할 상황이 옵니다.

이를 챗봇 형식으로 문제를 풀어나가보고자 Swagger Inspector를 개발하게 되었습니다.

프로젝트 기여 내용Codex 기반 개발 효율 증대

- Codex를 리팩터링, 디버깅, 반복 코드 정리에 활용
- 보조 도구로 사용하여 개발 속도를 높이고 작업 효율 개선
- 단순 반복 작업을 줄여 주요 기능 구현에 집중

RAG기반 신뢰도 있는 답변 출력

- MongoDB Vector Search & text-embedding-3-small을 통한 벡터화 · 임베딩 진행
 - 시스템 프롬프트에 OpenAPI 스펙을 넣지 않음으로서 토큰비용 절감
 - Threshold를 통해 사용자 입장에서의 신뢰도 추가

트러블슈팅 / 이슈사실과 다른 출력 응답상황

OpenAPI 관련 질문을 입력하였을 때, 질문과 유사하지 않은 기능들을 골라서 출력하는 상황이 발생했습니다.

이 때문에 사용자는 여러 기능들이 섞여있는 출력을 받아 혼란을 초래하였습니다.

해결방향

각 문서마다 질문에 대한 신뢰도를 도입하고, 유사한 문장을 검색할 때 신뢰도를 기반으로 탐색하도록 구조를 수정하였습니다.

또한 출력 시 어떤 문서를 참고했으며, 그 문서에 대한 신뢰도를 출력하는 방법으로 재설계 하였습니다.

지정된 신뢰도 임계치를 넘지 못하는 문서는 과감히 검색에서 제외하는 방향으로 수정하였고,

이 임계치는 사용자가 직접 조절할 수 있도록 변경하였습니다.

결과

사용자에게 믿을 수 있는 문서들을 추천하고, 그 근거들을 사용자에게 보여줌으로서 신뢰도가 높아졌습니다.

활동

입상내역

-  2025 교내 캡스톤 프로젝트 수상
 - AI가 추천해주는 청년정책 및 재정관리 서비스 - 은상
 - 온디바이스 LLM을 활용한 애완견 케어 서비스 - 동상
-  전국 고등학교 동아리 SW 경진대회
 - YOLO와 MediaPipe를 활용한 졸음운전 감지 서비스 - 장려
-  제 7회 데이터 크리에이터 캠프 - 장려
-  2025 SW-AI - 장려
-  2024 MS 해커그라운드 해커톤 - 장려

내·외부 활동

-  의성 자기소개페이지 워크샵 참가
-  Apple Foundation Program @ POSTECH 3기
-  2025 AI Expo Korea 국제인공지능대전 부스 참가 및 운영
-  2025~26학년도 선도부장
-  퀀텀아이 1개월 인턴십
-  2025년도 실리콘밸리 견학
-  2026 AI Expo Korea 국제인공지능대전 작품 출품 (Dogtor)
-  제3회 아진실리경진대회 작품 2개 출품 (압축공기 누설 및 절전 알림시스템, LightBox)

자격증

-  정보처리기능사 (2025 정기기능사 1회)
-  파이썬 프로그래밍 3급 (민간자격증, 한국인공지능아카데미 발급)
-  딥러닝활용자격증 3급 (민간자격증, 한국인공지능아카데미 발급)

etc.

프로젝트를 진행하며 모델을 설계하는 것보다, 실제 서비스 환경에서 문제를 마주하고 해결하는 과정이 더 재미있다는 것을 알게 되었습니다.

여러 트러블슈팅들을 해결하는 과정들이 저를 계속 개발로 끌어당겼습니다. 차후 AI 엔지니어링, 특히 모델을 안정적으로 서빙하고 운용하는 MLOps 방향으로 나아가고자 합니다.

이에 대해 공부하고 있는 내용들은 [블로그](#)와 [깃허브](#)에 기록중입니다.

긴 글 읽어주셔서 감사합니다. 인공지능소프트웨어과 최인규였습니다.