

강민석

AI 기술을 통해 일상의 문제를 해결하는 데 관심이 많습니다.

- 새로운 기술을 통한 창의적인 문제 해결에 강점이 있습니다.
- 부산대 FAQ 챗봇과 텍스트 임베딩 서버를 개발 및 배포했습니다.
- 📖 학력: 부산대 기계공학부 4학년 재학 (25-2 취업계 가능)
- 💻 GitHub: <https://github.com/myeolinmalchi>
- ✉ Email: minsuk4820@gmail.com

Tech Stack

- Backend & AI: FastAPI SQLAlchemy LangGraph LangChain
- Frontend: React.js TypeScript Zustand Next.js
- DevOps & Infra: Docker GitHub Actions EC2 ALB
- Database: PostgreSQL MySQL SQLite

🟡 : 최근 6개월 내에 주력으로 다룬 기술

🟢 : 최근에 사용하지 않았거나, 기본적인 이해를 갖춘 기술

주요 약력

외주/프리랜서 활동

- 22.04 – 22.05: Okra Seoul 게시판 서버 개발 (GoLang)
- 22.06 – 22.11: METAGONZ 랜딩페이지 제작 (Gatsby.js)
- 22.10 – 23.01: 천안아산강소특구 홈페이지 퍼블리싱 (HTML, CSS)
- 23.03 – 24.01: 한국기술허브 PoC 프로젝트 참여 (Next.js, FastAPI)

APPTIVE: 부산대 창업 동아리

- 24.03 - 24.06: 정규 스터디 활동 (프론트엔드)
- 24.09 – 24.12: 정규 프로젝트 활동 (프론트엔드 개발)
- 24.09 – 현재: 프론트엔드 멘토 (React.js 커리큘럼 제작 등)
- 25.03 – 현재: 금융 AI 비서 프로젝트 팀 (AI 백엔드)

AID: 부산대 인공지능 동아리

- 24.09 – 24.12: NLP 기초 스터디 (N-gram, BERT 등)
- 24.09 – 25.03: PNU Chat 프로젝트 개발 (프로젝트 총괄)
- 25.04 – 현재: AI 논문 리딩 스터디 (ReAct 등 논문 다수)
- 25.04 – 25.05: DAIC 2025 해커톤 운영 (Upstage 교육팀과 협업)
- 정규 세미나 발표 활동 (YOLO 추론 최적화, 프로젝트 이슈 트래킹 등)

PNU Chat: 부산대학교 학생지원 챗봇 서비스

개발 기간: 24.10 – 25.03 (재정비 및 팀원 추가 모집 중)



프로젝트 개요

학과 공지사항, 학생지원시스템 등의 데이터를 주기적으로 수집·가공하여 학생의 질의에 실시간으로 응답하는 **대화형 챗봇 서비스**를 구축 및 배포했습니다.

기능 고도화 및 정확도 개선을 위해 재정비 기간을 가지고, 팀원을 추가 모집 중입니다.

주요 경험 및 성과

- 공지 크롤링 및 임베딩 파이프라인 구축 → **24시간 내 15만 건** 데이터 처리
- GGUF 모델 도입 → CPU 추론 속도 **19.3% ↑**, 메모리 사용량 **71.3% ↓**
- Hybrid Score + RRF 전략 실험 → Q8 환경 **Recall@3 15.52%p ↑**
- 베타 버전 출시 직후, **일일 최대 200명** 수준의 트래픽을 무중단으로 처리

역할 및 기여

- 팀 구성: **총괄 1 (9할 이상 기여)**, 개발 보조 2, 디자이너 1
- 기획·개발·배포 전 과정 총괄, 역할 분배 및 일정 관리
- GitHub Flow 기반 협업 프로세스 정립 및 코드 리뷰 진행

PNU Chat: 부산대학교 학생지원 챗봇 서비스

개발 기간: 24.10 – 25.03 (재정비 및 팀원 추가 모집 중)

데이터 수집 & Retrieval 시스템 구축

- RDB Schema 설계 및 pgvector 기반 Hybrid Retrieval 구현
- DI 패턴 적용으로 Retrieval 전략과 비즈니스 로직 실험 효율화
- **15만 건**의 학과 홈페이지 공지사항 크롤링 파이프라인 구축

Multi-Agent RAG 구조 설계

- 기존 Fan-out/Fan-in 구조에서 과도한 토큰 소비 및 응답 지연 문제 분석
- 개선을 위해 LangGraph 기반 Supervisor 구조 설계 및 전환 계획
- 토큰 사용량과 지연을 줄이면서도 정확도를 향상하는 것이 목표

임베딩 최적화 & Recall 개선 실험

- 제한된 컴퓨팅 자원을 고려해 bge-m3 기반 임베딩 서버 구축
- GGUF Q8 양자화 적용 -> 추론 속도 19.3%↑, 메모리 71.3%↓
- 정확도 손실 개선을 위해 Hybrid Score + RRF 전략 설계 및 실험
→ Recall@3 Dense-only 대비 15.5%p↑, FP32 대비 하락폭 1%p 미만

[Infra] 인프라 구성 및 배포 자동화

- GitHub Actions 기반 CI/CD 파이프라인 구축 → EC2 배포 자동화
- EC2 + ELB + Route53 조합으로 스케일업 대비 및 커스텀 도메인 구성
- 웹 퍼블리싱 테스트 목적으로 Amplify를 통한 임시 배포 환경 운영

[FE] UI/UX 설계 및 구현

- 디자이너와의 협업을 통해 반응형 챗봇 UI/UX 설계
- 프론트엔드 전체 단독 개발, 로컬 저장소 기반 개인화 기능 구현
- 마크다운 응답의 하이퍼링크(첨부파일 포함)를 위한 커스텀 로직 구현

은행원 대상 금융 비서 (MVP)

개발 기간: 25.03 (3주, MVP 개발)

프로젝트 개요

은행 영업 직원의 업무 보조를 위한, 내·외부 문서 기반의 금융 챗봇 MVP입니다.
초기 MVP 대상은 상품이 다양하고 크롤링이 용이한 KB국민은행으로 선정했습니다.

역할 및 기여

- 팀 구성: 백엔드 1, 프론트엔드 1, 디자인 1, 기획 1
- Selenium 기반 금융 상품 크롤링 파이프라인 구축
- 금융 상품 RDB 스키마 설계 및 Hybrid Retrieval 통합
- 기존 PNU Chat 아키텍처 일부 재활용하여 구현 시간 단축



AI 금융상품 큐레이터

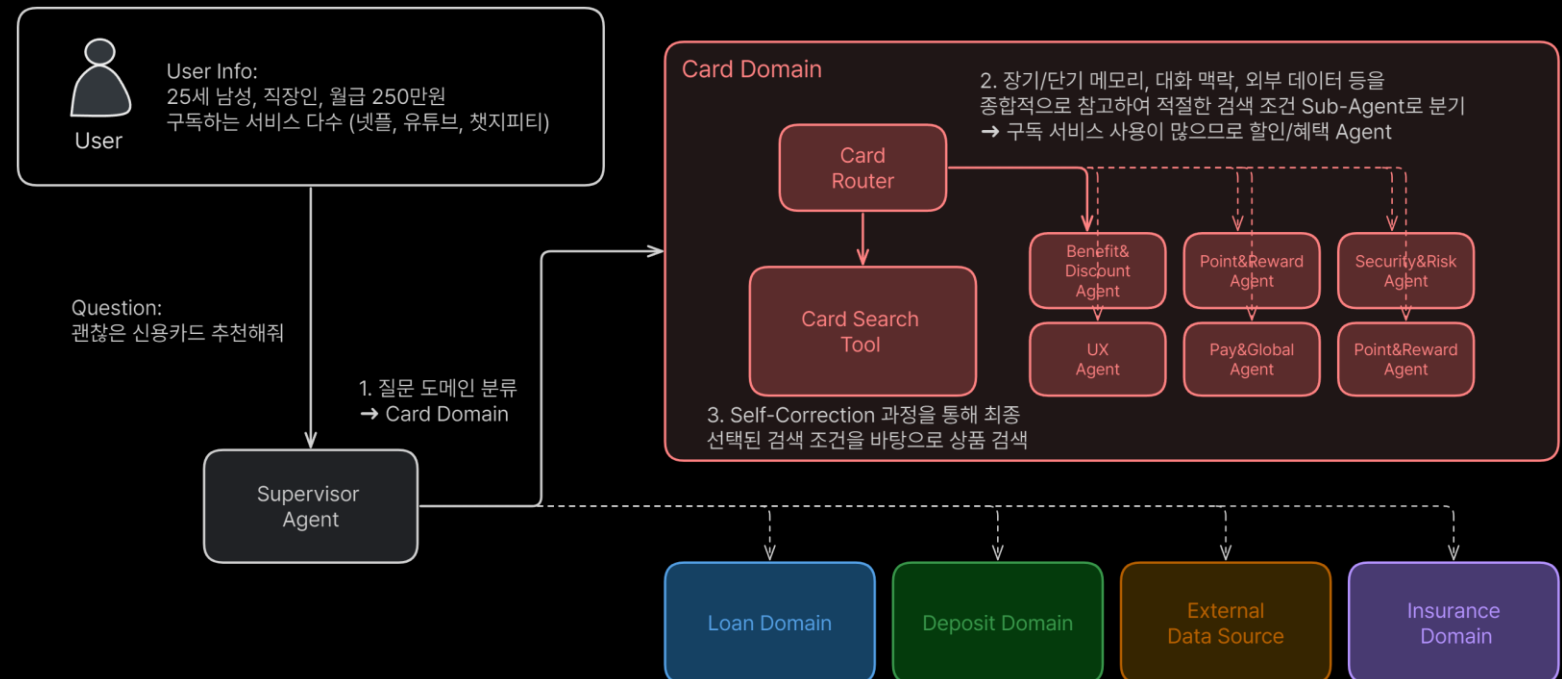
개발 기간: 25.05 – 현재 (기획 변경 후 PoC 단계)

프로젝트 개요

AI Agent가 사용자의 상황과 대화 맥락을 능동적으로 이해하고, 동적인 Scoring 전략을 수립하여 검색하는 **금융 상품 추천 시스템**입니다.

행동 기반 추천의 한계를 극복하기 위해, **기존에 없던 방식**을 시도합니다.

범용성을 고려하여 새롭게 기획하였으며, 현재 PoC 개발 단계입니다.



PNU x Upstage: DAIC 2025 해커톤 운영


활동 기간: 25.04 – 25.05 (개발지원팀)

Upstage 협업 & 대회 운영

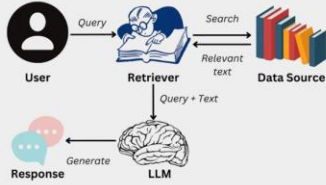
- Slack 채널을 통해 업스테이지 교육팀과의 긴밀한 협업
- 카카오톡 오픈채팅을 통해 참가자에게 대회 관련 공지 전달
- 중간점검 및 예선 심사를 위해 팀별 레포지토리 모니터링

OT(킵오프) 프레젠테이션 제작

PNU x Upstage
**DOCUMENT AI
CHALLENGE 2025**



관련 기술: RAG



관련 기술: in-context learning

Zero-shot Prompting
다음 문장의 강령을 분석하세요:
"나 오늘 우물에서 뽕 샀어."

One-shot Prompting
예시: "오늘은 너무 행복한 하루였어" → 긍정
예시를 참고하여 다음 문장의 강령을 분석하세요:
"나 오늘 우물에서 뽕 샀어."

Few-shot Prompting
예시1: "이런, 트윈프 때문에 또 마이네이션!" → 부정
예시2: "중간고사는 망했지만 이번 초연대였어!" → 중립
예시3: "오늘은 너무 행복한 하루였어!" → 긍정
예시를 참고하여 다음 문장의 강령을 분석하세요:
"나 오늘 우물에서 뽕 샀어."

주관: AI Developer 후원: Upstage

주제 관련 기술 소개 문서 작성

AI Agent (Agentic Workflow)

AI Agent는 단순한 질의응답 서비스를 넘어서, 능동적으로 목표를 설정하고, 계획을 수립하며, 주어진 태스크를 수행할 수 있는 인공지능 시스템을 의미합니다. Agent는 추론과 계획 그리고 기억이 가능하며, 자율성을 갖고 능동적으로 의사 결정, 응답 품질 평가 등을 수행할 수 있습니다.

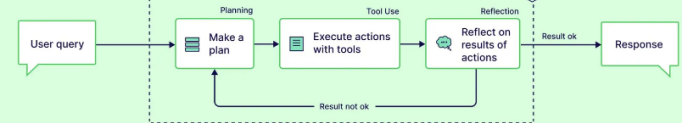
Automated Workflow (rule-based, non-AI)



AI Workflow (non-agentic)



Agentic Workflow



AI Agent는 복잡한 작업을 수행하기 위해 다음과 같은 **Agentic Workflow**를 따릅니다.

- **계획 세우기**: LLM은 복잡한 작업을 더 작은 단위의 하위 작업으로 나눈 다음, 우선 순위를 정합니다.
- **도구 사용**: Agent는 각 하위 작업을 수행하기 위해 미리 정의된 도구를 상황에 맞게 선택 및 사용합니다. 가령, "오늘 날씨 어때?"라는 질문에 대해, 사용자의 위치 정보를 기반으로 실시간 날씨 정보를 불러오는 API를 도구로써 사용할 수 있습니다.
- **중간 점검과 재시도**: Agent는 각 하위 작업을 수행해 나가면서, 중간 결과의 품질을 점검하고, 만족스러운 결과가 나올 때까지 재시도 합니다. 필요한 경우 전체 계획을 조정할 수도 있습니다.

AI Agent는 외부 세계와 소통하고, 여러 Agent와 정보를 주고받으며 자신의 임무를 수행합니다. 이를 통합하는 시스템을 설계하고, 다양한 구조를 실험하면서 보다 유연하고 지능적인 AI 서비스를 구현할 수 있습니다.

APPTIVE 프론트엔드 멘토

활동 기간: 24.09 – 현재 / 커리큘럼 및 과제물 제작

컴포넌트 순수하게 유지하기

함수형 컴포넌트의 설계 철학에 따르면, 렌더링 로직은 순수해야 합니다.

다시 말해,

- 동일한 입력에 대해 항상 동일한 JSX를 반환하고,
- 렌더링 과정에서 외부에 영향을 끼치지 않아야 합니다.

동일 입력 동일 출력

함수형 컴포넌트의 입력은 `props` 와 `state (context)`입니다.

`state` 는 컴포넌트 내부의 상태값으로, `useState` 를 통해 주입됩니다.

이는 렌더링 로직 바깥에서만 변경이 허용되며, 컴포넌트 관점에서 입력으로 취급될 수 있습니다.

그러므로, 함수형 컴포넌트는 동일한 `props` 및 `state` 에 대해 항상 동일한 JSX를 렌더링 해야합니다.

부수 효과

앞서 언급했듯이, `setter` 함수를 호출하여 `state` 를 변경하는 것 자체는 부수 효과에 해당합니다.

React는 이러한 `setter` 함수 뿐만 아니라, 데이터 fetching, DOM 조작 등으로 발생하는 모든 부수 효과를 렌더링 로직으로부터 분리하는 API를 제공하는데, 그것이 바로 `useEffect` 입니다.

```
import { useEffect } from 'react'

const Component = () => {
  // ...
  const [count, setCount] = useState(0)
  useEffect(() => {
    console.log('저는 부수 효과입니다.');
```

}, [count]);

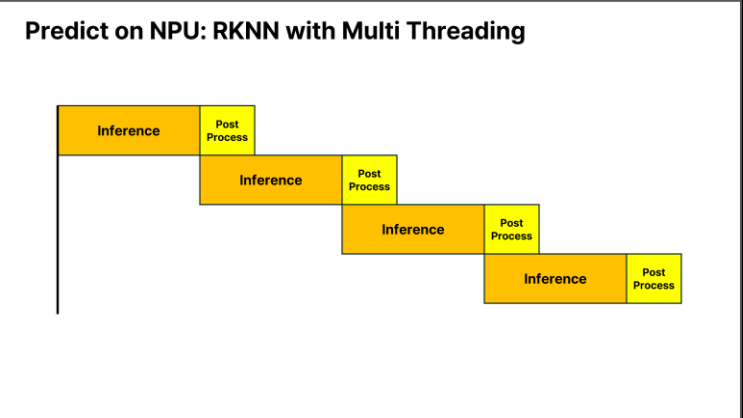
```
  // ...
  return <>안녕하세요</>
};
```

`useEffect` 의 첫 번째 인자로 전달되는 콜백 함수를 `Effect` 라고 하며, 이는 렌더링 결과물이 실제 DOM에 적용 된 뒤에 실행됩니다. `Effect` 에 부수 효과를 일으키는 로직을 포함함으로써 여전히 렌더링 과정을 순수하게 유지할 수 있습니다.

- 문서화 및 커리큘럼 부실 문제 개선을 위해 멘토로 자원
- React.js 공식 문서를 바탕으로 학습자료 및 과제 제작
- 실습 위주의 방식으로 전환, 과제 이행을 약 50% 개선

세미나: 엣지 디바이스에서의 YOLO 추론 성능 최적화

대상: 부산대 AID / 일시: 2024.12.27



- NPU 기반 SBC에서의 YOLO 추론 최적화 경험 공유

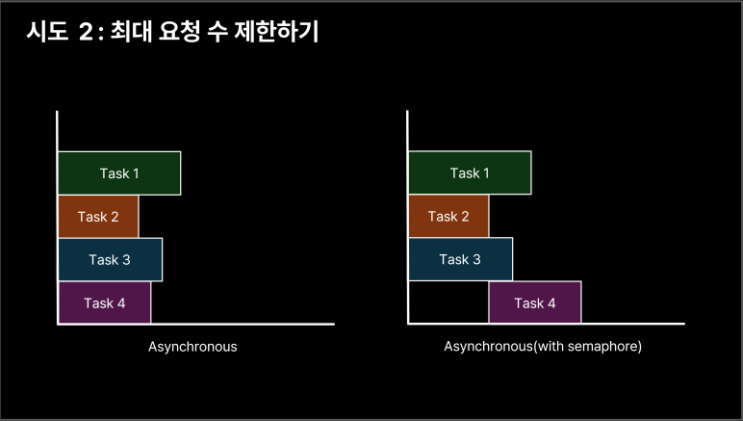
세미나: 은행 내부 직원 전용 금융 비서 AI 챗봇 (MVP)

대상: 부산대 APPTIVE / 일시: 2025.04.04



세미나: PNU Chat 프로젝트

대상: 부산대 AID / 일시: 2025.04.04



- 개발 과정에서의 이슈 트래킹 경험 공유

AID 논문 리딩 스터디

활동 기간: 25.04 – 현재

1. Introduction

Chain-of-Thought(CoT)의 한계:

- 모델의 추론 과정이 “내부 세계”에만 의존
- “외부 세계”와는 단절되므로 Knowledge Update가 불가능
- Hallucination 및 Error propagation 문제

>> Reasoning 및 Acting을 결합하여 General task solving에 적용 시도

(1) Botspot QA

Question: Aside from the Apple Remote, what other device can control the program Apple Remote was originally designed to interact with?

(1a) Standard

Answer: iPod

(1b) CoT (Reason Only)

Thought: Let's think step by step. Apple Remote was originally designed to interact with Apple TV. Apple TV can be controlled by iPhone, iPad, and iPod Touch. So the answer is iPhone, iPad, and iPod Touch.

Answer: iPhone, iPad, iPod Touch

(1c) Act-Only

Act 1: Search[Apple Remote]

Obs 1: The Apple Remote is a remote control ...

Act 2: Search[Front Row]

Obs 2: Could not find [Front Row]. Siriarr ...

Act 3: Search[Front Row Software]

Obs 3: Front Row is a discontinued media center software ...

Act 4: Finish[OK]

- ReAct, ReflAct 등 AI Agent 관련 논문 리딩 및 발표