

류지수

AI Engineer | LLM, RAG, AI Agent

LLM 기반 AI 시스템의 문제 정의부터 설계, 평가, 서비스 적용까지 경험한 AI Engineer입니다. RAG·Tool Calling 기반 AI Agent와 Text2SQL 시스템을 개발하고, PII 탐지 모델의 Knowledge Distillation으로 추론 지연을 개선했습니다. 문제 유형에 맞는 검색·실행 구조를 선택하고, 실패 가능성을 고려해 설계해왔습니다.

Experience

GoyoAI / AI Engineer

2023.12 - 현재

역할 기반 AI Agent 구조 설계 및 개발

AI Agent / Tool Calling / RAG

문서 검색, 업무 데이터 조회, 외부 검색을 Tool로 구성해 질문에 맞는 기능을 Agent가 선택하도록 설계했습니다.

- 질문 → Tool 선택 → 데이터 조회 → 결과 가공 → 답변 생성 흐름 구현 (문서 검색은 RAG·Reranker, 업무 데이터는 백엔드 API 호출)
- 직전 2턴의 대화 이력을 함께 전달해 후속 질문의 맥락 유지
- 판단과 응답 생성은 LLM이, 데이터 조회는 Tool이 맡도록 역할을 분리하고 조회 결과는 LLM이 처리하기 적합한 형태로 가공
- Tool 호출이 20회를 넘으면 실행을 중단하도록 제한해 반복 호출로 인한 무한 루프 방지
- 백엔드, 기획 담당자와 Tool 스펙과 데이터 범위를 협의해 현업에 적용

LangChain Tool Calling RAG Reranker

외부 LLM용 PII 보안 계층 및 경량 탐지 모델

PII Detection / Knowledge Distillation

Qwen3 8B 기반 PII 탐지를 KLUE RoBERTa-large로 증류해 테스트셋 평균 inference latency를 8.35초 → 0.06초로 단축했습니다.

- 외부 LLM 호출 시 사내 개인정보가 유출될 위험을 줄이기 위해, 기획 요구사항에 따라 호출 전 입력과 문서의 PII를 탐지해 차단하고 관리자가 허용한 경우에만 마스킹 후 호출하는 정책 기반 보안 계층 구현
- 추론 지연을 줄이면서 탐지 품질을 유지하는 것을 목표로 Qwen3 8B를 Teacher, KLUE RoBERTa-large를 Student로 두고 OCR 문서와 synthetic data로 학습 데이터를 구성해 Knowledge Distillation 적용
- OCR 특유의 개행과 문자 변형을 반영한 약 50개 케이스로 Teacher와 Student의 탐지 결과 및 오류 유형을 비교한 결과, 생년월일, 계좌번호, 주소, 성별 등 주요 Entity에서 Teacher와 대부분 일치하는 탐지 결과 확인
- 한자 낱자 표기, 성별 오탐, 주소 범위 과확장 케이스를 한계로 확인하고 개선 과제로 도출

Qwen3 8B KLUE RoBERTa-large Knowledge Distillation

자연어 기반 정형 데이터 질의 (Text2SQL)

Text2SQL / Tool Calling / PostgreSQL

Vector Search 기반 RAG의 한계를 해결하기 위해 자연어 질의를 SQL로 변환해 PostgreSQL에서 직접 실행하는 구조로 전환하고 운영 환경에 적용했습니다.

- 정형 데이터 질의에서 조건에 맞는 데이터가 여러 건이어도 유사도 검색 특성상 일부 Chunk만 검색되어, 검색이 아닌 직접 조회가 맞다고 판단
- pgvector로 유사 질문-쿼리 64개를 검색하고 bge-reranker-v2-m3로 상위 5개 Few-shot을 선별, 테이블 DDL, 컬럼 설명, 샘플 3행을 Schema Context로 제공
- SQL 실행 실패 시 DB 오류 메시지를 LLM에 전달해 수정 후 재실행 (최대 3회)
- 최종 SQL을 재실행해 전체 결과를 Interactive Table과 CSV, Excel Export로 제공

pgvector bge-reranker-v2-m3 PostgreSQL

Technical Skills

LLM Systems	RAG, Reranker, Tool Calling, AI Agent, Text2SQL, LangChain, LangGraph (PoC)
Model	Knowledge Distillation, PII Detection (Qwen3 8B, KLUE RoBERTa-large)
Backend / Data	Python, FastAPI, PostgreSQL, pgvector

Education

울산대학교 IT융합 / 소프트웨어공학	2018.03 – 2023.02
Naver Boostcamp AI Tech 4기 추천 시스템 (Recsys)	2022.09 – 2023.02

Certificates

ADsP 데이터분석 준전문가 한국데이터산업진흥원	2024.06
-------------------------------	---------