SK 네트웍스 Family AI 과정 14기 개발된 LLM 연동 웹 애플리케이션 모델배포

개발된 LLM 연동 웹 애플리케이션

제출일자 : 2025.09.29. (금)

작성 팀원 : 이승혁

| 개요 | **On-premise AI 기반 신입 행원 고객응대 지원 웹앱 구축.**  **추진 목표**   1. 신입 행원의 **고객 응대 품질·속도** 향상 및 **업무 표준화** 2. KB 고유 지식의 **체계적 축적·활용**과 **보안 요건 충족(온프레미스)** 3. 여신 심사 업무의 **정확성·효율성 제고**와 보고 자동화   **주요 기능** A. 지식형 챗봇  * **KB 특화/일반 금융 Q&A** 제공 (사내 문서 기반 RAG, sLLM 파인튜닝 적용) * **근거 제시**: 참조 **PDF 문서 및 페이지 색인**을 응답에 명시 * **FAQ 6종 고정 노출**로 빠른 셀프가이드 제공 * **채팅 기록 라이브러리**: 대화 이력 저장· 검색·재활용 * **문서 리포지토리 관리**: 현재 저장 문서 현황/버전 관리  B. 여신 심사 지원  * **개인·기업 여신 심사** 모듈 (머신러닝 모델 기반) * **자동 보고서 생성·확인**(심사 결과 요약, 근거, 지표 시각화 포함)  C. 사내 커뮤니케이션  * **사내 메신저** 연동/대화 기능 * **공지·중요 정보 게시/열람** 화면 제공  기술·운영 원칙  * **온프레미스 운영**으로 고객/사내 데이터 외부 반출 방지 * **RAG + sLLM 파인튜닝**을 통한 정확도·도메인 적합성 확보 * **감사 로그/접근 통제** 및 문서 출처 추적(페이지 단위)로 **컴플라이언스 강화**  기대 효과  * 신입 중심의 **응대 리드타임 단축** 및 **답변 일관성** 확보 * 여신 심사 **의사결정 지원 고도화**(근거 기반, 보고 자동화) * 사내 정보 접근성 개선으로 **업무 생산성 향상**   **기술 스택**   1. 프론트엔드 : html, CSS, JavaScript, Django 2. 백엔드: FastAPI, Django 3. DB: MySQL, Pinecone 4. 배포: AWS Elastic Beanstalk, EC2, RDS, S3, Route 53, vllm 5. 기타: OpenAI, PDFPlumber, Runpod |
| --- | --- |
|  | **프로젝트 모델 실행 순서**  KB FinAIssist 프로젝트는 Django 웹 애플리케이션과 FastAPI AI 추론 서버로   구성됩니다   1. **백엔드 FastAPI 실행**   SLM 기반 Langgraph 및 RAG(Retrieval Augmented Generation) 기능을 담당하는 AI 추론 FastAPI를 실행합니다.   1. **프론트엔드 Django 실행**   사용자 인터페이스와 비즈니스 로직을 담당하는 Django 웹 서버를 실행합니다. 이 서버는 FastAPI 서버와 통신합니다.   1. **모델 실행**   Hugging-face로부터 파인튜닝된 Qwen2.5-7B-Instruct 모델을 로드하고, Runpod 환경에서 langgraph와 연동하여 추론을 수행합니다. 모델에 활용된 query, output, 참조된 pdf는 모두 mysql에 연동 및 저장됩니다. 또한 S3 연동을 통해 참조된 pdf의 전문을 플랫폼에서 바로 확인할 수 있습니다.  **필요한 패키지**  각 환경별로 필요한 패키지들을 설치합니다.   1. **Backend 환경 (FastAPI)**   pip install -r requirements.txt   | # Core API  fastapi==0.116.1  uvicorn==0.35.0  python-dotenv==1.1.1  python-multipart==0.0.20  anyio==4.10.0  # LangChain & workflow  langchain-core==0.3.75  langchain-openai==0.3.31  langchain-huggingface==0.3.1  langchain-pinecone==0.2.11  langchain-text-splitters==0.3.9  langgraph==0.2.67  # Vector store & model clients  pinecone==7.3.0  openai==1.101.0  # Document ingestion  pandas==2.3.2  pdfplumber==0.11.7  tqdm==4.67.1  # Evaluation helpers  requests==2.32.5 | | --- |  1. **Django 환경**   pip install -r requirements.txt   | # requirements.txt (웹앱 전용)  Django==5.2.5  gunicorn==23.0.0  asgiref==3.9.1  sqlparse==0.5.3  mysqlclient==2.2.7  pymysql==1.1.1  pandas==2.3.2  numpy==2.3.2  scikit-learn==1.7.0  boto3==1.34.142  python-dotenv==1.1.1  requests==2.32.5  django-extensions==4.1  django-ebhealthcheck  uvicorn  whitenoise  # 데이터 시각화  plotly==5.17.0  dash==2.14.2  dash-bootstrap-components==1.5.0 | | --- | |

- 1 -

SK 네트웍스 Family AI 과정 14기 개발된 LLM 연동 웹 애플리케이션 모델배포

| 기본 사용법 | **홈페이지 사용법**  **1. main 화면 - 로그인 페이지(redirect)**  a. 사원일 경우 → 로그인  b. 관리자일 경우 → 관리자 로그인  **2. 메인 dashboard 화면**  a. KB 업무 지원 챗봇 선택  ⅰ. 질의응답  ⅱ. 사내 pdf 문서 실시간 참조하여 대답 생성  ⅲ. 내규에 따른 답변 필터링(가드레일)  b. 개인/기업 여신 심사 선택  ⅰ. 머신러닝 모델 기반 여신 심사 보고서 제공  **3. KB 업무 지원 챗봇 화면**  a. sLLM 모델과 채팅  ⅰ. 일반 금융 지식 질의응답  ⅱ. 자주 묻는 질문 목록 -> 빠른 질의응답  ⅲ. 새로운 채팅 시작  ⅳ. 채팅 검색 및 채팅 히스토리 확인  ⅴ. 참조된 pdf 확인 가능  **4. 여신심사 화면**  a. 개인/기업 여신 심사 토글  ⅰ. 고객 정보 입력 -> DB 내 고객 정보 확인  ⅱ. 대출 정보 입력 -> 여신 심사 요청  ⅲ. 머신러닝 기반 여신 신용 평가 및 승인 여부 예측  ⅳ. 신용 분석 보고서 제공  ⅴ. 보고서 저장 및 불러오기 |
| --- | --- |
| 확장 및  커스터마이징 | **1. 시스템 분야 확장성**   1. 메인페이지에 등록된 기능 활성화   ⅰ. KB 업무 지원 챗봇 (O)  ⅱ. 개인/기업 여신심사 (O)  ⅲ. 금융상품 조회 (TBD)  ⅳ. 개인/기업 고객 조회 (TBD)  ⅴ. 보고서 관리 (TBD)  ⅵ. 일정 관리 (TBD)  ⅶ. 사내 보안 메신저 (TBD) |

- 2 -

SK 네트웍스 Family AI 과정 14기 개발된 LLM 연동 웹 애플리케이션 모델배포

| 결론 | **성과**  **fine-tuned sLLM 기반 AI 웹 애플리케이션 구축**   * KB 도메인 문서에 최적화된 **소형 LLM(sLLM) 파인튜닝** 및 **RAG** 결합으로 정확도와 응답 속도 확보   **FastAPI(추론) × Django(프런트) 연동 아키텍처 완성**   * Django UI/업무흐름 ↔ FastAPI 모델서빙(질의·랭킹·근거생성) **비동기 연계**   **AWS Elastic Beanstalk PoC 배포**   * 운영·확장성 검증 완료(PoC). **최종 목표는 온프레미스 이전**(보안/규제 준수)   **금융·KB 특화 데이터 학습 sLLM로 질의응답 검증**   * **PDF 근거 페이지 인용**(페이지 색인)과 **FAQ 6종 고정 노출**로 신입 행원 셀프가이드 강화   **머신러닝 기반 개인/기업 여신심사 + XAI**   * **설명가능성(XAI)** 포함: 중요 특성/지표, 규정 근거, 참고 문서 **즉시 확인** 가능   **신입(IT/개발 직군 포함) 창구업무 부담 경감**   * 표준화된 답변 템플릿, 근거 첨부, 자동 보고서로 **교육·헬프데스크 의존도 감소**     **향후 발전 방향**   1. 더 큰 모델과 더 많은 데이터로 챗봇 퀄리티 향상 2. 실제 기업 의뢰 시, 내부 보안 데이터 접근으로 실전 업무 배치 가능 3. 데이터 및 디자인 변경을 통한 타 금융기관으로의 확장 가능 |
| --- | --- |

- 3 -