SK네트웍스 Family AI 과정 14기

프로젝트 기획서

산출물 단계	기획
평가 산출물	프로젝트 기획서
제출 일자	2025.10.01
깃허브 경로	SKN14-Final-1Team
작성 팀원	김준기, 김재우, 안윤지, 이나경, 이원지희, 정민영

1. 프로젝트 주제

LLM 활용 내부 고객 업무 효율성 향상을 위한 API 전문 개발자 지원 AI 기반 문서 검색 시스템

2. 문제 정의

API 문서는 다양한 서비스와 기능을 제공하지만, 그 자료와 문서가 방대하고 구조가 복잡해 신속하고 정확한 검색이 쉽지 않습니다. 내부 고객 및 개발자들은 필요한 정보를 찾기 위해 많은 시간을 소비하고 있으며, 업무 생산성 저하로 이어지고 있습니다. 이로 인해 실무에서는 문서 이해도 부족, 관련 자료 누락, 비효율적인 검색 과정 등의 문제가 빈번히 발생하고 있습니다.

따라서 내부 업무 효율성을 높이기 위해, RAG + LLM 기반의 전문화된 문서 검색 및 지원 시스템이 필요합니다. 저희 프로젝트는 개발자가 필요한 정보를 빠르게 찾을 수 있도록 지원하며, 업무 과정에서 발생하는 질의와 요구에 효과적으로 대응할 수 있도록 API 챗봇 AI(외부 API 문서 지원용)와 기업 내부 문서 sLLM(사내 전용 챗봇)을 함께 제공할 예정입니다.

이 시스템은 단순한 정보 검색을 넘어, 개발 환경 전반에서 의사결정을 신속하게 지원하고 프로젝트 진행 속도를 높이도록 도움을 줄 것입니다.

3. 시장조사 및 BM

3-1. 시장 조사

a. 시장 규모 및 성장 추이

- **글로벌 AI 개발자 지원 시장: 2028년 시장 규모 6,788억 2,000만 달러 전망**

2028년까지 글로벌 AI 소프트웨어 시장 규모가 6,788억 2,000만 달러에 도달할 것으로 예상됩니다. 이는 공인된 보고서에 기반한 것으로, 인공지능 소프트웨어 시장이 급성장하고 있다는 여러 지표들을 명확히 보여줍니다. 이러한 성장은 특히 산업 전반에서 AI 활용의 확대, 엣지 AI와 IoT의 결합, 헬스케어 분야에서의 AI 채택 증가 등이 주요 원인으로 작용할 것입니다.
(출처: 글로벌 AI 개발자 지원 시장)

- **국내 시장**

과학기술정보통신부와 소프트웨어정책연구소의 '2024 인공지능산업실태조사'에 따르면, 2024년 국내 인공지능(AI) 산업은 6조 3,000억 원에 달하는 것으로 나타났습니다. AI 응용 소프트웨어(챗봇, 자동화 등)가 2조 6,700억 원으로 전체 AI 사업 매출에서 가장 큰 비중을 차지했습니다. 이어 △AI 구축·관리 및 정보 서비스(컨설팅, 클라우드 등) 1조 8,700억 원 △AI 시스템 소프트웨어(머신러닝 플랫폼, 추론 엔진 등) 1조 4,600억 원 △AI 연산 처리 부품·장치(NPU 등) 3,000억 원 순이었습니니다. 매출 증가율 역시 응용 SW(14.3%), 구축·관리 서비스(13.1%), 시스템 SW(9.6%), 연산 부품(8.4%) 순으로, 응용 서비스 분야의 고성장세가 두드러졌습니다.
(출처: http://www.itdaily.kr/news/articleView.html?idxno=233950)

b. 타겟 고객(Target Audience)

- 외부 API를 활용하는 내부 개발자 및 엔지니어
- API 관련 문서 탐색과 활용이 중요한 기술 지원팀 및 기획자
- API 서비스 및 연동 솔루션을 개발하는 사내 관련 부서

3-2. BM

[가치제안 (Value Proposition)]

- LLM 기반 AI 문서 검색으로 구글API 및 관련 방대한 자료를 신속·정확하게 탐색하여 필요한 정보를 즉시 확보
- 개발자가 쉽게 API를 이해할 수 있도록 예제 코드, 사용 가이드, 주요 파라미터 설명 등 맥락 기반 맞춤형 해설 제공
- 기업 내부 전용 sLLM을 통해 내부 문서/자료를 안전하게 검색하고, 보안 규정에 부합하는 답변 제공
- 개발자 맞춤형 실시간 질의응답으로 API 학습 속도를 높이고, 프로젝트 개발 효율을 극대화
- 내부 개발자 및 고객의 업무 효율성과 결과물 품질을 동시에 향상
- 보안 필터와 접근 제어를 통해 기밀 데이터 유출 방지 및 안정적인 사내 협업 환경 조성

고객	기업 개발팀, 스타트업, 시스템 통합(SI) 업체, 클라우드 서비스 파트너사
수익 모델	- 구독형 라이선스: 월/연 단위 과금, 사용자 수·기능별 차등 요금제 - 엔터프라이즈 계약: 기업 맞춤형 데이터 연동·배포, 기술 지원 포함
제공 가치	- 외부 API 및 사내 문서 기반 Q&A, 코드 예시·오류 해결 자동화 - 대화내역 카드 저장·검색 기능은 사용자가 마이페이지에 대화 기록을 체계적으로 보관하고 필요 시 신속하게 검색하여 활용할 수 있도록 지원하는 개인 맞춤형 레퍼런스
차별성	- API 문서 기반 챗봇으로, 오류 원인부터 코드 예시·사용법까지 전 과정을 정확하고 단계적으로 안내 - API 문서를 검색하면 해당 페이지의 URL이 제공되어, 원하는 API 문서 페이지로 바로 접속 가능능 - 로컬 SLLM으로 기업 내부 문서 검색 기능 제공 가능 → 보안 요구 충족 - 파인튜닝을 통한 공손/친구말투 지원

4. 시스템 구성 기획

본 시스템은 Django/Fastapi 기반 웹 애플리케이션과 여러 기능 모듈로 구성되며, AWS 환경에서 Docker 컨테이너로 배포됩니다.

전체 아키텍처는 백엔드(Django-Gunicorn/Uvicorn, FastAPI), 벡터 DB(챗봇 RAG 데이터 검색용), MySQL(사용자 및 콘텐츠 저장)로 이루어져 있습니다. MySQL은 AWS EC2 인스턴스 내부에서 운영되며, Django 애플리케이션과 Fastapi 애플리케이션은 각각각 Elastic Beanstalk 환경에 Docker 컨테이너 형태로 배포됩니다.

API 챗봇 서비스는 (구글) API 사용법 Q&A 코드 예시 제공, 오류 해결, 관련 질문 추천, 대화 내역 카드 저장 기능 등을 지원합니다. API 챗봇은 OpenAI LLM을 활용하여 Django 내에서 서비스되며, 사내 내부 문서 챗봇은 QWEN 모델을 기반으로 FastAPI 기반 비동기 API 서버와 RUNPOD(VLLM SERVERLESS)를 활용하여 동작합니다. 커뮤니티 기능에서는 게시글에 '좋아요' 수가 5개 이상이면 자동으로 베스트 게시글로 선정됩니다.

또한 API 문서 검색 엔진은 검색어에 대한 유사도 기반으로 관련 API 문서 링크 리스트를 제공합니다. 인증·승인 시스템은 회원가입 시 직급 정보를 입력받고, 관리자가 승인해야 계정이 활성화되는 방식으로 운영됩니다.

백엔드 서버는 Django에서는 Gunicorn과 Uvicorn을 함께 사용하여, Django 전통 요청(동기 처리)과 챗봇·API 호출(비동기 처리)을 병행함으로써 성능을 최적화합니다.

특히 FastAPI는 비동기 방식으로 RunPod VLLM 서버에 추론 요청을 보내고 응답을 받아 Django 백엔드로 전달함으로써, 고성능 LLM 연동과 서비스 응답 속도를 동시에 확보합니다.

5. 모델링 계획

[핵심 자원]

- LLM 모델 + RAG(벡터 DB)
- API 문서 수집·임베딩 파이프라인
- AWS + Docker 기반 서버 인프라
- sLLM 파인튜닝 모델 + RAG(벡터 DB)

[API 챗봇(외부 개발 지식 전문 → 우선 구글 API부터 지원)]
: API 개발 관련 질문에 바로 답해주는 전문가형 챗봇

- API 챗봇 AI는 우선 구글 API 문서(11개) 개발과 관련된 질문에 즉시 답변할 수 있도록 설계된 전문가형 시스템입니다. 이를 위해 구글 API 공식 문서(txt)를 적절한 단위로 분할하여 임베딩하고 벡터 DB에 저장합니다. 또한 수집된 문서를 기반으로 QA 데이터셋을 구축해 함께

벡터 DB에 보관합니다.

- 사용자가 질문을 입력하면 시스템은 원문 벡터 DB와 QA 벡터 DB를 유사도 기반으로 검색하여 관련된 정보를 찾아냅니다. 이후 LLM은 검색된 문서를 바탕으로 답변을 생성하며, 프롬프트 설계를 통해 답변의 품질을 높입니다다. 추가로 연관 질문도 함께 제안합니다. 만약 생성된 답변이 부정확하거나 불완전하다고 판단될 경우, 시스템은 문서를 다시 검색하고 LLM을 다시 호출하여 최종 답변을 다시 보여주는 방식으로 사용자가 더욱 정확하고 신뢰할 수 있는 정보를 얻을 수 있도록 합니다.

-

[기업 내부 문서 sLLM(내부 비서 + 보안 필터 → 사내 전용 챗봇)]
: 조직 내부 규정, 정책, 기술 자료를 검색하고 권한에 맞는 답변 제공

- 프론트엔드·백엔드·AI·데이터팀·CTO 직급/팀 내부 문서를 GPT 프롬프트를 활용해 Markdown 형식의 텍스트 파일로 작성하고, 각 문서에 (백엔드/AI, 데이터/프론트엔드팀, CTO 등)과 같은 권한 정보를 메타데이터 태그로 부여합니다. 문서는 청킹한 뒤 임베딩되어 권한 메타데이터(role)와 함께 벡터 DB에 저장합니다. 사용자가 로그인 후 채팅을 하면 DJango가 해당 사용자의 권한을 확인 후, 중앙 게이트웨이인 FastAPI 서버에 채팅 메세지와 사용자의 권한을 전달합니다. Fastapi 서버에서는 Django에서 받은 채팅 메세지와 팀·역할 정보를 확인하고, 해당되는 권한 범위로 연결된 벡터 DB에서 유사도 기반 문서 검색을 수행합니다.. FastAPI는 검색된 상위 청크를 수집해 RAG용 프롬프트를 구성하고, 필요 시 사내 시스템에 대한 툴 호출을 중개해 추가 사실(문서 검색 내용)을 조회한 뒤, 컨텍스트와 함께 sLLM에 전달하여 최종 답변을 생성하게 한다. 이때 모든 질의·검색·툴 호출·모델 응답은 모니터링과 품질 개선을 위해 로깅되며, 실패·지연 등 예외도 FastAPI 레벨에서 일관되게 처리됩니다. 또한 공손말투/친구말투 스타일을 반영하도록 sLLM을 파인튜닝하여, 친구 말투의 기본 예시는 “안녕 나는 ***이야”, 공손 말투의 기본 예시는 “안녕하세요 저는 ***입니다” 와 같이 정의하여 사용자가 원하는 말투로 sllm 챗봇이 답변할 수 있도록 합니다.

6. 사용 데이터

- **구글 API 공식 문서**
  - Google for developers
  - API 별 링크
    - OAuth 2.0 / Google Identity
    - People API
    - Google Drive API
    - Google Sheets API
    - Gmail API
    - API Reference
    - YouTube Data API

        - Google Maps
        - Firestore REST API
        - Firebase Authentication
        - BigQuery API

- 사내 내부 문서

  백엔드팀 문서(15개)

  데이터/AI팀문서(15개)

  프론트엔드팀 문서(15개)

  CTO 문서(15개)

7. 역할분담(R&R)

- 총괄 : 김준기
- 데이터수집 :
    - 김준기 : OAuth 2.0 / Google Identity, YouTube Data API
    - 이원지희 : People API
      이나경 : Drive API, Sheets API
    - 안윤지 : Gmail API, Calendar API
    - 정민영 : Firebase/Firestore API, Firebase Authentication
    - 김재우 : BigQuery API,Maps Platform
- AI 개발(백엔드)
    - API문서 RAG 기반 LLM 개발 : 안윤지, 정민영
    - API 문서 챗봇 RAG평가 :이나경,안윤지,정민영
    - 사내 문서 SLLM 파인튜닝 : 김준기, 정민영
    - 사내 문서 SLLM 성능 평가 : 정민영
    - 사내 문서 SLLM - FASTAPI/RUNPOD 연동 : 김준기, 정민영
- django 프론트엔드
    - 로그인/회원가입: 이원지희
    - 승인대기/거부페이지:이원지희
    - 마이페이지:안윤지
    - API 챗봇페이지:김재우,이나경
    - API 문서검색:이나경
    - 사내내부문서 챗봇페이지:김재우
- django 백엔드
    - 로그인/회원가입:이원지희

    - 승인거부페이지:이원지희
    - 커뮤니티페이지:김재우
    - 사내내부문서페이지:김준기, 정민영
    - API 챗봇페이지:정민영,이나경
    - API문서검색:이나경
    - 마이페이지:안윤지
- 서비스 배포 : 김준기, 정민영
    - 서버/인프라 담당 (AWS 환경 구축·네트워크·스토리지 설정)
    - 앱 컨테이너 & 웹 서버 담당 (Django/Gunicorn-Uvicorn 및 FATAPI Docker화 및 Nginx 설정)
    - 배포 & 모니터링 담당 (깃허브 ACTION기반 CI/CD 구축 및 배포 후 성능·로그 점검)