

**SK네트웍스 Family AI 과정 13기  
 데이터 전처리 학습된 인공지능 모델**



| **산출물 단계** | 데이터 전처리 |
| --- | --- |
| **평가 산출물** | 학습된 인공지능 모델 |
| **제출 일자** | 2025.08.10 |
| **깃허브 경로** | <https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN13-FINAL-6Team> |
| **작성 팀원** | 최성장, 김동욱, 안수민 |

1. **모델 목적**: 사내 임직원들을 위한 업무 도우미 서비스
2. **모델 아키텍처 설계**

* 선정 모델: gpt-4o를 사용한 RAG 체인
* 아키텍처 개요:

| **계층** | **구성 요소** | **역할** |
| --- | --- | --- |
| 문서 검색 | Qdrant, nlpai-lab/KoE5 | 코사인 유사도 기반으로 질의와 유사한  문서 청크 검색 이후 context 생성 |
| 응답 생성 | gpt-4o | 질의와 context를 이용한 답변 생성 |

* 아키텍처 시각화:
  + 사용자 입력 → RAG 문서 검색 → 관련 문서 Context 생성 → GPT-4o 응답 생성 → 후처리 및 사용자 전달
* 설계 근거
  + nlpai-lab/KoE5는 intfloat/multilingual-e5-large의 한계를 보완하기 위해 선정함. 임베딩할 문서는 한국어로 구성되어 있으므로, 한국어 특화 파인튜닝된 해당 모델 다국어 모델(intfloat/multilingual-e5-large) 대비 한국어 문서 검색에 대해 성능 및 비용 균형이 우수할 것으로 판단함.
  + 오픈소스 모델로 라이센스의 제약이 적음
  + Gpt-4o는 정확도 및 속도 면에서 성능이 우수했음.

1. **모델 학습 요약**

* 추가 학습 없음(Fine-tuning미적용)
* 문서 검색의 정확도는 임베딩 모델 성능에 의존.
* GPT-4o는 사전 학습된 상태 그대로 API호출

1. **저장 및 배포**

* 문서 저장소 : Qdrant(VectorDB)
  + 한국어 문서 청크 단위 벡터 저장
  + 메타데이터 포함(문서명, 카테고리, 부서)
* 배포환경
  + 벡터 DB : Qdrant, EC2(Docker 컨테이너)
  + PostgreSQL(RDB) : Amazon RDS

1. **종합 평가 및 활용 방안**

* 장점
  + 사내 문서 기반 맞춤형 답변 제공
  + API기반으로 유연한 확장 가능
* 활용 방안
  + FAQ자동 응답 시스템
  + 사내 규정 및 정책 검색 서비스

1. **추가 기재**

* 향후 개선 방향
  + 대화형 메모리(Conversation RAG) 도입
  + 한국어 LLM 파인튜닝 적용 가능성 검토
  + 사내 일정/ 이메일/ 프로젝트 관리 툴과 통합