(주)에셈블_정수장 공정 맞춤형 챗봇 개발

1. 프로젝트 개요

본 프로젝트는 정수장 공정 운영자의 질의에 대해 **정확하고 빠르게** 맞춤형 답변을 제공하는 챗봇 시스템을 개발하는 것을 목표로 한다.

챗봇은 오픈소스 LLM과 SLM을 기반으로 구축하며, PDF 등 다양한 문서에서 추출한 도메인 데이터를 활용하여 현장 운영에 특화된 자연어 응답을 제공한다.

---

2. 개발 범위 (챗봇 파트)

본 개발자는 전체 시스템 중 **챗봇 엔진 및 데이터 처리 파트**를 전담하며, 다음 기능을 구현한다.

1. **질문 해석 모듈 (LLM 기반)**
   - 사용자의 자연어 질문을 해석하고 의도를 파악.
   - 질의 내용에서 핵심 요구사항 추출 (예: "침전지 이상 소독 처리량" → `침전지`, `소독 처리량`).
2. **데이터 매칭 모듈**
   - 사전 구축된 **정수장 공정 데이터베이스**에서 관련 문서/절차/규정 검색.
   - 해당 데이터베이스는 PDF, 매뉴얼, 보고서 등에서 추출된 문서를 분석하여 **단락별 주제·태그**가 부여된 형태로 저장.
   - 빠른 검색을 위해 벡터 DB(Faiss, Milvus 등) 활용.
3. **응답 생성 모듈 (LLM 기반)**
   - 검색된 데이터와 사용자 질문을 기반으로 자연어 응답 생성.
   - 전문 용어 유지 + 사용자가 이해하기 쉬운 설명 제공.
   - 불필요한 정보 제거 및 핵심 요약.

---

3. 시스템 구조

```
사용자 질문 → [질문 해석 LLM] → 요구사항 추출
          ↓
      [데이터 검색 모듈] → 벡터DB에서 관련 데이터 로드
          ↓
     [응답 생성 LLM] → 자연어 최종 답변 반환
```

---

4. 데이터 준비

- **데이터 출처**
  - 정수장 매뉴얼, 설계 보고서, 유지보수 지침, 법규 문서, 기술 보고서.
- **전처리 방식**
  - PDF/Word/Excel → 텍스트 변환 (pdfminer, unstructured, PyMuPDF)
  - 문단별 주제 분류, 키워드 태깅
  - 벡터화 (Sentence-BERT, KoSimCSE) → 벡터 DB 저장
- **검색 최적화**
  - 유사도 검색 + 키워드 필터링
  - 문서 메타데이터(공정명, 설비명, 규격 등) 기반 필터

---

5. 기술 스택 및 모델

오픈소스 LLM 후보 (정확도 & 한국어 지원)

모델명	특징	장점	단점
KoAlpaca / KoVicuna (LLaMA 기반)	한국어 특화 파인튜닝	한국어 질문 해석 및 응답 품질 우수	경량 버전 속도 제한
Polyglot-Ko (3.8B / 5.8B)	대규모 한국어 데이터 학습	속도·정확도 균형	5.8B는 GPU 메모리 요구 높음
OpenKo-LLM	한국어 QA에 특화	최신 LLaMA2 기반, 높은 정확도	상대적으로 배포 초기 단계
MPT-7B-Instruct (한국어 파인튜닝)	빠른 응답 속도	처리 속도 빠름	한국어 데이터셋 부족 시 품질 저하 가능

SLM 후보 (질문 의도 분석, 라우팅 등 경량 처리)

모델명	용도	장점
KoMiniLM	질문 의도 분석, 키워드 추출	가볍고 빠름
KoSimCSE	벡터 검색을 위한 문장 임베딩	한국어 유사도 검색 최적
DistilKoBERT	짧은 텍스트 분류	GPU 자원 절약

---

6. 기대 효과

- 현장 운영자가 **매뉴얼 검색 없이** 필요한 정보를 즉시 획득.
- 긴급 상황 시 빠른 대응 가능 (예: 수질 이상 발생 시 대처 방법).
- 신규 직원 교육 자료로 활용 가능.