**1. 사용자 화면**

**1.1 그룹 공통 Chat UI**

* 기능: 기본 채팅 기능, 사용자 프로필 표시, 최근 대화 기록, 알림 기능.
* 기술 스택: React, Vuㄴe.js
* 추가 제안: 최신 상태 관리 라이브러리인 Recoil을 사용해 상태 관리를 최적화할 수 있습니다.

#### 1.2 업무 기능별 Chat UI

* **R&D**
  + 기능: 기술 문서 검색, 연구 보고서 요약, 특허 검색.
  + 기술 스택:
    - **프론트엔드**: React, Vue.js
    - **백엔드**: Node.js, Express
    - **데이터베이스**: **PostgreSQL**, MongoDB
    - **검색 기능**: Elasticsearch, Apache Solr
    - **NLP 모델**: SpaCy, BERT, Hugging Face Transformers
* **영업**
  + 기능: 고객 관리, 판매 보고서 생성, 제품 정보 제공.
  + 기술 스택:
    - **프론트엔드**: React, Vue.js
    - **백엔드**: Node.js, Express
    - **데이터베이스**: **PostgreSQL**, MySQL
    - **CRM 통합**: Custom CRM API
    - **데이터 시각화**: D3.js, Chart.js
* **구매**
  + 기능: 공급업체 정보, 주문 상태 추적, 재고 관리.
  + 기술 스택:
    - **프론트엔드**: React, Vue.js
    - **백엔드**: Node.js, Express
    - **데이터베이스**: **PostgreSQL**, MySQL
    - **ERP 시스템 통합**: Custom ERP API
    - **재고 관리**: Redis, Apache Kafka
* **재무**
  + 기능: 재무 보고서 생성, 예산 추적, 비용 분석.
  + 기술 스택:
    - **프론트엔드**: React, Vue.js
    - **백엔드**: Node.js, Express
    - **데이터베이스**: **PostgreSQL**, MySQL
    - **재무 데이터 통합**: Custom Finance API
    - **분석 도구**: Pandas, NumPy
* **분류, 요약, 번역**
  + 기능: 문서 분류, 요약 생성, 실시간 번역.
  + 기술 스택:
    - **프론트엔드**: React, Vue.js
    - **백엔드**: Node.js, Express
    - **NLP 모델**: BERT, RoBERTa, T5, MarianMT
    - **데이터베이스**: **PostgreSQL**, MongoDB
    - **번역 API**: MarianMT
    - **분류 : NLP 모델 외** Scikit-learn, TensorFlow, PyTorch 등

**2. 질의 분석**

**질의 언어 식별**

* 모델: FastText, langid.py

**구조 분석**

* 모델: SyntaxNet, SpaCy

**의미 분석**

* 모델: BERT, RoBERTa, ALBERT

**질의 구체화**

* 모델: BERT, RoBERTa, ALBERT

**개인화**

* 데이터: 직무, 소속, 역할, 검색 이력.
* 기술 스택:
  + **사용자 프로필 데이터베이스**: PostgreSQL, MongoDB
  + **세션 관리**: Redis
  + **사용자 인식 및 관리 시스템**: Keycloak, Auth0

**3. Vector DB 외 Storage**

**Vector DB(필드 메타)**

* 기술: Milvus

**Vector DB(RAG)**

* 기술: FAISS

**질의문 DB**

* 기술: PostgreSQL, MongoDB

**참고문서(Doc)**

* 기술: Elasticsearch

**질의 Log DB**

* 기술: ELK Stack (Elasticsearch, Logstash, Kibana)

**응답 Log DB**

* 기술: ELK Stack

**4. 결과 처리**

**RAG 기반 답변**

* 모델: Retrieval-Augmented Generation 모델 (BERT + GPT), RAG-Token

**응답 생성**

* 모델: GPT-4, T5, GPT-4 Turbo

**응답 검증(AI Filter)**

* 기술: Rule-based 필터, 추가 검증 모델 (e.g., BERT), RoBERTa

**5. LLM 모델**

* **하이퍼클로바**
  + 특징: 초대형 한국어 모델, 다양한 도메인 커버리지.
* **Gemini**
  + 특징: 다국어 지원, 고성능 NLP.
* **sLLM: Llama3**
  + 특징: 오픈소스 LLM, 높은 확장성.
* **sLLM: Mistral**
  + 특징: 경량화된 LLM, 빠른 응답 속도.
* **추가 제안:**
  + Claude 2 (Anthropic)
  + Meta의 최신 LLaMA 시리즈 모델
  + Falcon

**6. 관리도구**

**LLM Version 관리**

* 기술: Git, Model Registry (e.g., MLflow)

**학습 Data 관리**

* 기술: 데이터 레이크 (e.g., Delta Lake)
* 추가 제안: Apache Hudi

**질의/답변 Log**

* 기술: ELK Stack, BigQuery
* 추가 제안: Datadog

**사용자 Feedback**

* 기능: 피드백 수집, 분석, 반영.
* 기술: 설문조사 도구, 대시보드.
* 추가 제안: Typeform

**Prompting Template**

* 기능: 프롬프트 템플릿 저장, 관리.
* 기술: 템플릿 관리 시스템, YAML/JSON 형식 파일.
* 추가 제안: Prefect