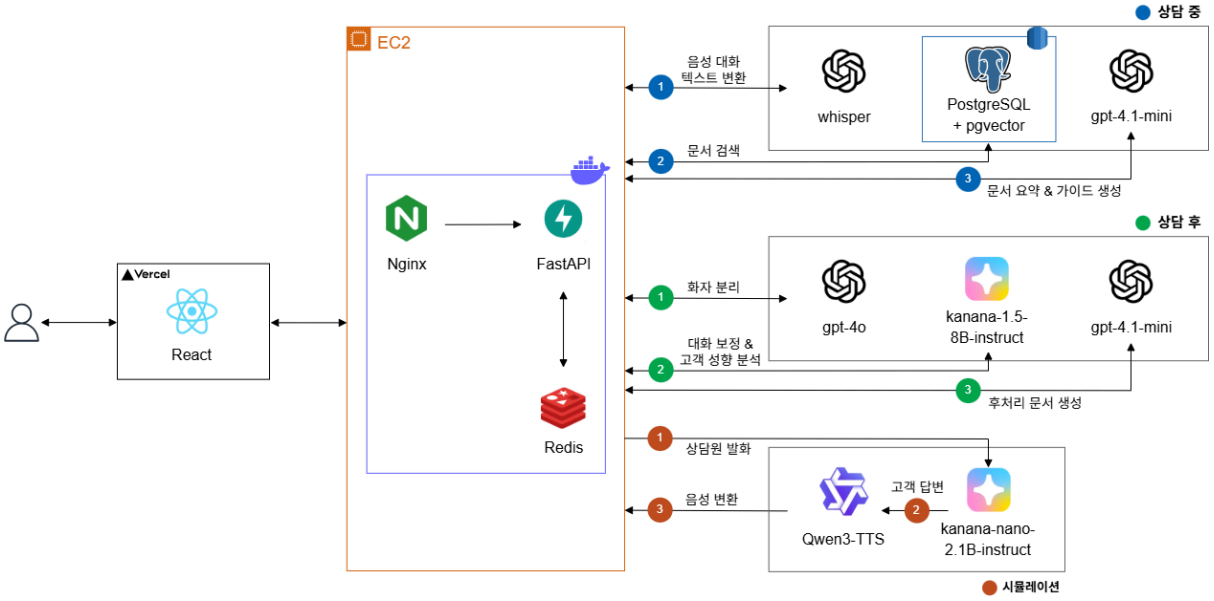
시스템 아키텍처



1. Frontend

- **React**

  사용자로부터 음성 입력을 받고, 백엔드와 통신

---

2. Backend

- **Nginx**

  프록시 서버로서 외부의 요청을 받아 내부 도커 컨테이너로 전달

- **FastAPI**

  시스템의 메인 로직을 담당

  프론트엔드와의 WebSocket 연결을 유지

  STT, RAG, TTS 기능 오케스트레이션

- **Redis**

  캐싱 담당

---

3. 핵심 서비스 모듈

- **상담 중**
  1. STT
     a. FastAPI로부터 전달받은 음성 데이터를 OpenAI Whisper 모델 엔드포인트로 전송
     b. 음성 데이터가 텍스트로 변환되어 백엔드 → 프론트엔드로 반환
  2. Vector DB: PostgreSQL(+pgvector) 데이터베이스에서 유사한 문서 검색
  3. 검색된 문서의 요약, 전문, 상담 가이드 멘트 생성 `OpenAI gpt-4.1`

- **상담 후**
  1. 상담 전문 화자 분리(고객/상담원) `OpenAI gpt-4o`
  2. STT로 생성한 대화 전문 보정 `kanana-8b-1.5-instruct`
  3. 고객의 발화에서 고객 성향 파악 - 파인튜닝한 `kanana-8b-1.5-instruct` 모델 사용
  4. 후처리 문서 생성 `OpenAI gpt-4.1`

- **시뮬레이션**
  1. 상담원 발화를 상담 중 흐름과 동일하게 STT로 변환
  2. 변환한 텍스트로 상담 답변 생성 - 파인튜닝한 `kanana-nano-2.1b-instruct` 모델 사용
  3. 생성된 텍스트를 `Qwen3-TTS`로 전달하여 음성 파일로 변환
  4. 생성된 음성을 백엔드 → 프론트엔드로 반환