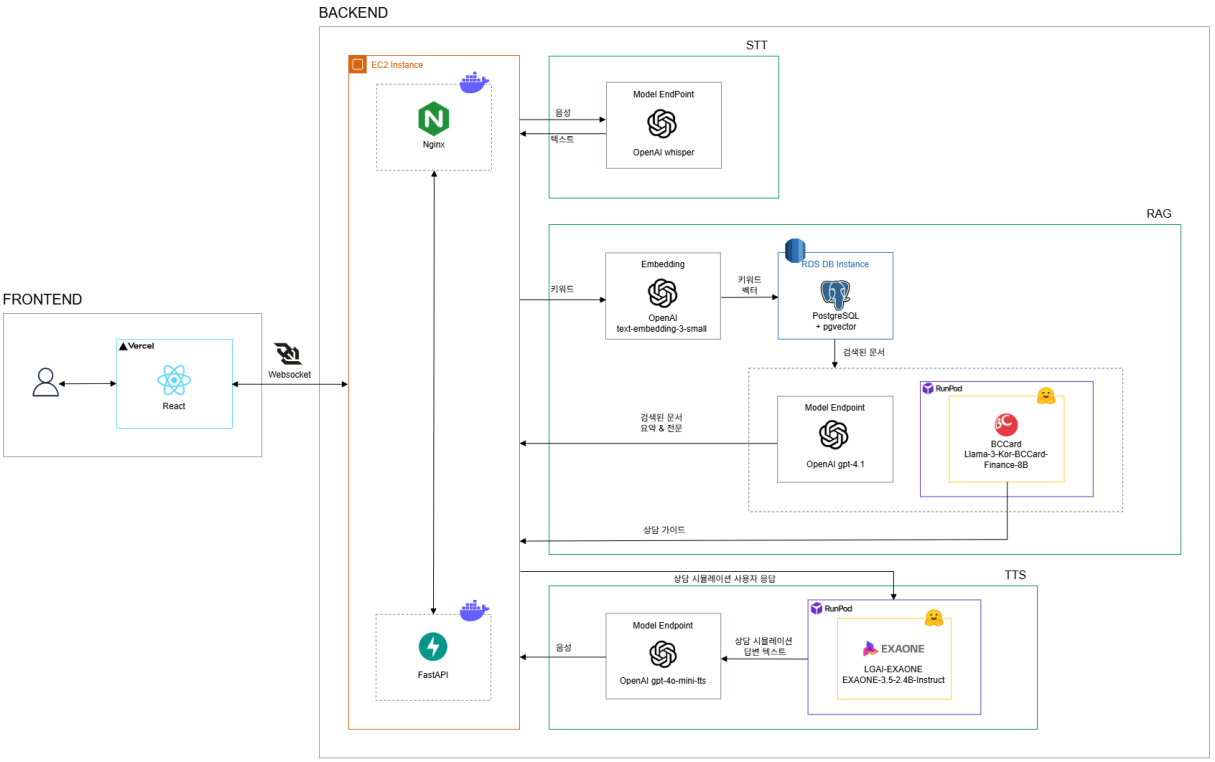
시스템 아키텍처



1. Frontend

- **React (Vercel을 통해 배포)**

  사용자로부터 음성 입력을 받고, 백엔드와 WebSocket을 통해 실시간으로 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

---

3. 핵심 서비스 모듈

- **STT**
  1. FastAPI로부터 전달받은 음성 데이터를 OpenAI Whisper 모델 엔드포인트로 전송
  2. 음성 데이터가 텍스트로 변환되어 백엔드 → 프론트엔드로 반환

- **RAG**
  1. Embedding: `OpenAI text-embedding-3-small` 모델을 사용하여 키워드를 벡터로 변환
  2. Vector DB: PostgreSQL(+pgvector) 데이터베이스에서 유사한 문서 검색
  3. LLM
     - `OpenAI gpt-4.1` : 검색된 문서의 요약 및 전문 생성
     - `Llama-3-Kor-BCCard-Finance-8B` : 상담 가이드 생성
  4. 생성한 데이터를 백엔드 → 프론트엔드로 반환

- **TTS**
  1. 상담 답변 생성: `EXAONE-3.5-2.4B-Instruct` 모델을 통해 상담원의 응답에 적절한 답변 텍스트 생성
  2. 생성된 텍스트를 **OpenAI gpt-4o-mini-tts**로 전달하여 음성 파일로 변환
  3. 생성된 음성을 백엔드 → 프론트엔드로 반환