LotoAI System Architecture

System Overview

LotoAI is a modular AI services platform built with a microservices architecture. The system consists of three main components: Gateway (FastAPI), Agent Orchestrator, and RAG Server. Each component communicates via REST APIs and shares data through PostgreSQL and Qdrant vector database.

Architecture Components

Gateway Component: The Gateway serves as the main entry point for all client requests. It runs on port 8088 and handles routing to appropriate backend services. The Gateway implements rate limiting, authentication, and request validation.

Agent Orchestrator: This component manages AI interactions using OpenAI's API. It runs on port 8090 and provides chat functionality with conversation history tracking. The orchestrator includes a fallback stub mode for development without API keys.

RAG Server: The Retrieval-Augmented Generation server handles document upload, indexing, and semantic search. It uses Qdrant for vector storage and supports both OpenAI and local Sentence Transformers for embeddings. The server implements advanced features like reranking, multi-query retrieval, and unlimited chunking.

Technical Specifications

Database: PostgreSQL 15 stores metadata, user sessions, and upload records. Vector Database: Qdrant 1.7 manages embedding vectors with HNSW indexing. Message Queue: NATS handles asynchronous communication between services. Object Storage: MinIO provides S3-compatible storage for uploaded files.