BOT GPT – Conversational Backend Design Document

Overview

BOT GPT is a backend platform for multi-turn conversational AI supporting Open Chat and RAG modes.

Architecture

API Layer (Django Rest Framework) orchestrates requests, Service Layer handles RAG, summarization and LLM calls, Database persists conversations, messages and documents, External LLM performs reasoning.

Tech Stack

Python, Django Rest Framework, SQLite/PostgreSQL, Groq LLaMA models, Docker, GitHub Actions.

Database Design

Core entities include Conversation, Message, Document, DocumentChunk and ConversationDocument.

RAG Flow

Uploaded documents are chunked, relevant chunks are retrieved per query and injected into the LLM prompt as system context.

Context Management

Summarization is triggered when token limits are exceeded to control cost and maintain long conversations.

Scalability

Stateless APIs, database-backed memory, and model-agnostic LLM integration enable horizontal scaling.