Intellect Chat: Enterprise RAG Platform for Intelligent Multi-Agent Conversations

Overview

Intellect Chat is a sophisticated, self-hosted Retrieval-Augmented Generation (RAG) platform engineered to power intelligent, context-aware, and secure conversational AI applications. Designed for enterprises requiring deep customization and data sovereignty, it leverages a multi-agent architecture built on state-of-the-art open-source LLMs and vector retrieval technology. The platform seamlessly integrates diverse knowledge sources shared, agent-specific, and user-provided to deliver highly accurate, personalized responses. By automating complex query resolution and ensuring all data processing remains on-premise, Intellect Chat enhances user engagement, operational efficiency, and strict compliance with a strong emphasis on privacy and custom agent-based workflows

Solution & Implementation

1. Architected a scalable, multi-agent RAG system using gRPC and Docker to handle complex query routing, parallel knowledge retrieval, and context-aware response generation, all within a secure on-premise environment.
2. Engineered a sophisticated vector processing pipeline that chunks, embeds, and stores diverse data types (documents, images, audio) into a high-performance vector database for ultra-fast semantic similarity search.
3. Implemented a dynamic agent selection framework, featuring a supervisor agent for intelligent query routing and specialized agents for domain-specific tasks.
4. Developed advanced prompt engineering and composition techniques to optimally combine retrieved context with user queries.

TechStack

- Python
- gRPC (Google Remote Procedure Call)
- Llama.cpp & other Open-source LLMs
- Vector Embedding Models
- Qdrant(Vector Database)
- MongoDB
- Docker
- FastAPI

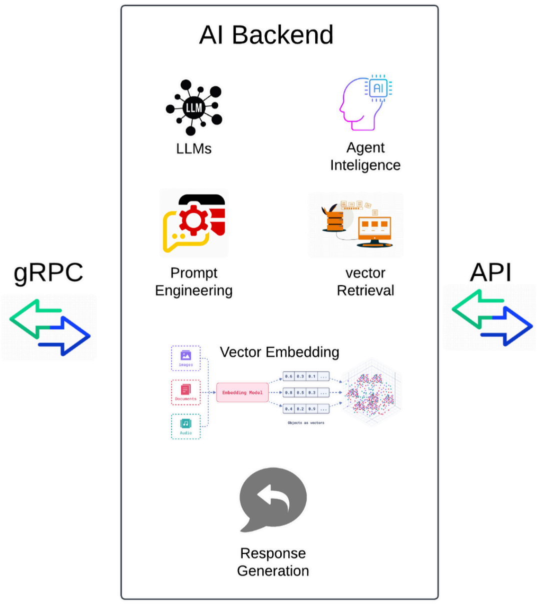
Customer Profile

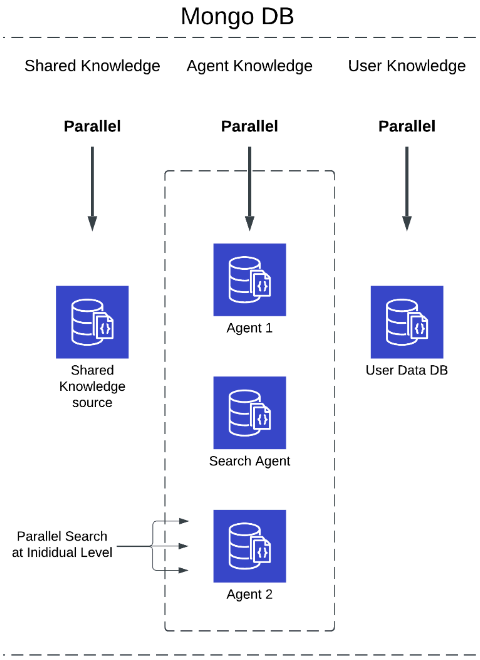
A discerning enterprise client requiring a secure, intelligent knowledge management
Industry: Enterprise Technology
Data Policy: Strict On-Premise

Results

The Intellect Chat platform was successfully developed and deployed as a custom RAG solution, providing the client with a powerful and private conversational AI engine. The system demonstrates exceptional performance in processing complex queries across multiple knowledge domains, significantly reducing response time and improving answer accuracy. The client now benefits from a unified platform that automates customer support, internal knowledge retrieval, and data analysis, all while maintaining full control over their sensitive data within their own infrastructure. This project showcases deep expertise in building full-stack, multi-modal AI systems with a strong emphasis on privacy and custom agent-based workflows.

Application Flow


AI Backend
LLM
LLMs
AI
Agent Inteligence
Prompt Engineering
vector Retrieval
Vector Embedding
Embedding Model
Response Generation
gRPC
API


Mongo DB
Shared Knowledge
Agent Knowledge
User Knowledge
Parallel
Parallel
Parallel
Agent 1
Shared Knowledge source
User Data DB
Search Agent
Parallel Search at Inididual Level
Agent 2