Agentic RAG Chatbot Architecture
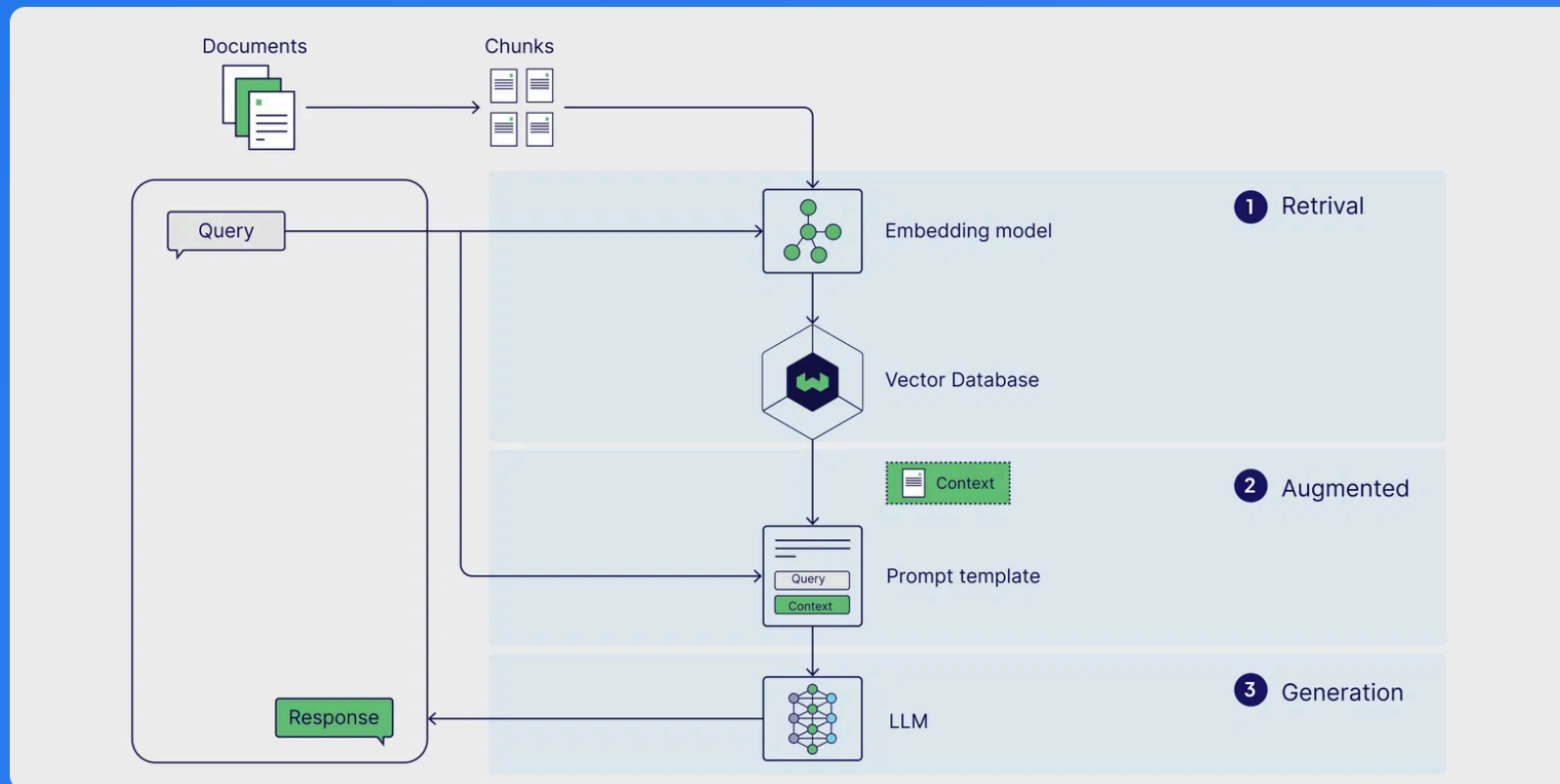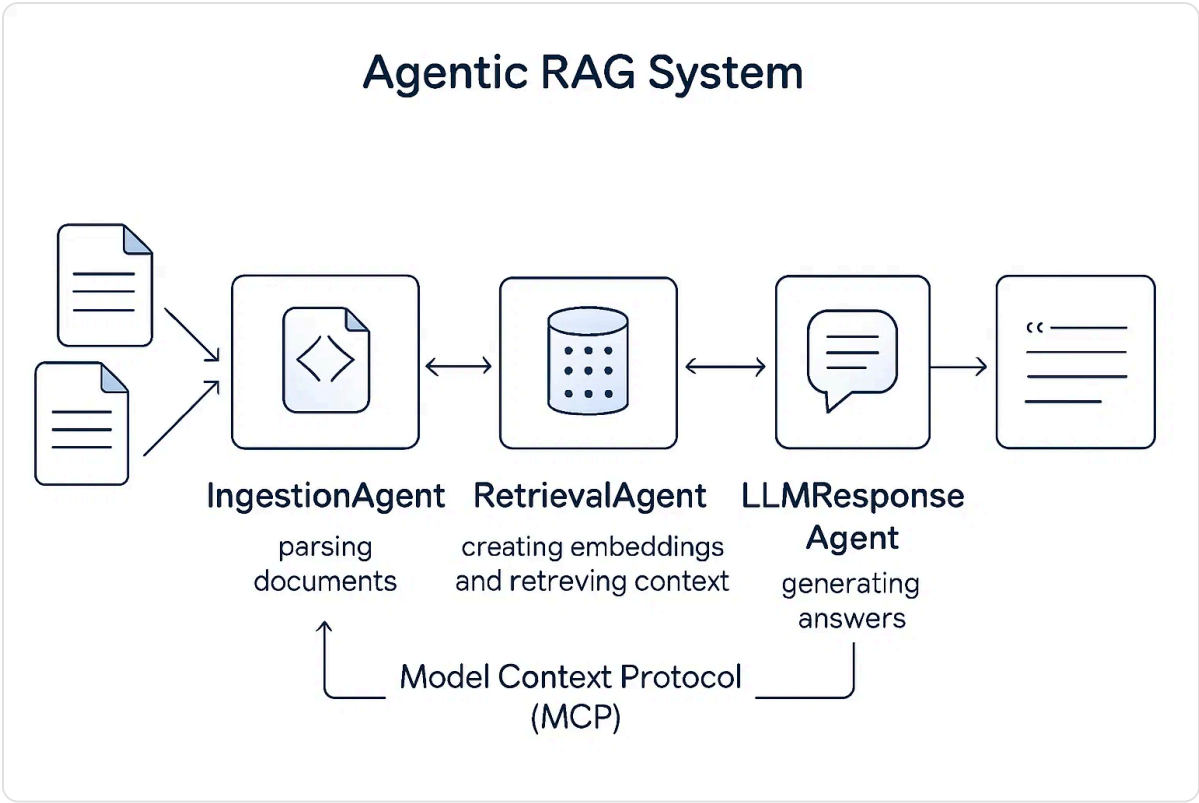
A Multi-Agent System for Document Question Answering

Agent Architecture

Agentic RAG System

IngestionAgent — parsing documents

RetrievalAgent — creating embeddings and retreving context

LLMResponse Agent — generating answers

Model Context Protocol (MCP)

⇥ IngestionAgent

Handles document parsing, chunking, and preprocessing. Supports multiple file formats and prepares content for embedding.

🔍 RetrievalAgent

Creates vector embeddings using Sentence Transformers and performs similarity search with FAISS to find relevant context.

💬 LLMResponseAgent

Generates final answers using Google Gemini model based on the retrieved context and user query.

Model Context Protocol (MCP): A structured messaging system that enables agents to communicate with each other through standardized message formats. MCP ensures clear data flow and traceability throughout the system.