RAG Financial Chatbot with n8n + OpenAI + Pinecone + Google Drive

Overview

This project demonstrates an **end-to-end Retrieval-Augmented Generation (RAG)** pipeline built using **n8n**, enabling a chatbot that can answer financial or business-related questions based on uploaded documents.
The workflow connects **Google Drive**, **OpenAI**, and **Pinecone** to create an automated system that processes, stores, and retrieves contextually relevant information to support natural language queries.

Workflow Logic

1. Google Drive to Pinecone (Data Ingestion & Indexing)

- **Trigger:** Detects new financial documents uploaded to Google Drive.
- **Download:** The file is automatically fetched using the Google Drive API.
- **Text Extraction & Splitting:** The document is parsed and divided into smaller chunks using the **Recursive Character Text Splitter**.
- **Embedding Generation:** Each chunk is converted into high-dimensional vectors using **OpenAI Embeddings**.
- **Vector Storage:** The embeddings are stored in **Pinecone**, enabling fast semantic search and retrieval.

2. Chat Interface (RAG Query Workflow)

- **Trigger:** Activated when a chat message is received.
- **Retrieval:** Relevant document chunks are fetched from **Pinecone Vector Store** based on the user's query.
- **AI Response Generation:** The **OpenAI Chat Model** combines retrieved context with the user's question to generate a grounded, conversational response.
- **Memory:** A simple memory module retains recent context for multi-turn interactions.

Tech Stack

- **n8n** – Workflow orchestration and automation engine
- **OpenAI API** – For embeddings and chat completion
- **Pinecone** – Vector database for semantic retrieval
- **Google Drive** – Source of uploaded financial or business documents

---

Use Cases

- Financial report Q&A automation
- Enterprise knowledge assistants
- Document-based advisory bots for investment, insurance, or compliance
- Custom internal chatbot for organization-specific policies and data

---

Setup Summary

1. Configure **Google Drive**, **OpenAI**, and **Pinecone** credentials in n8n.
2. Connect the **Google Drive Trigger** to automatically detect new uploads.
3. Add nodes for **Default Data Loader**, **Recursive Text Splitter**, and **OpenAI Embeddings** to prepare and vectorize document data.
4. Store embeddings in **Pinecone** using the **Pinecone Vector Store** node.
5. Build a second workflow that uses an **AI Agent** with **OpenAI Chat Model** and **Pinecone Retriever** to generate responses.

Outcome

Every time a new document is added to Google Drive, it's automatically indexed and made queryable. When users interact with the chatbot, it retrieves and summarizes relevant content, providing accurate, **data-grounded responses** in real time.

This workflow showcases how low-code tools like **n8n** can be combined with **OpenAI** and **Pinecone** to deploy production-grade RAG systems with minimal code.