RAG-BASED CHATBOT WITH GOOGLE GEMINI & PINECONE IN N8N

1. Overview

This workflow implements a **Retrieval-Augmented Generation (RAG) chatbot** in **n8n**.
It allows users to upload documents (via Google Drive), automatically **embed** them into Pinecone using **Google Gemini embeddings**, and then query these documents in real time with an **AI Agent powered by Google Gemini Chat Model**.

2. Workflow Architecture

The workflow consists of two main pipelines:

A. Document Ingestion Pipeline

1. **Google Drive Trigger**
    - Watches for new files uploaded to a specific folder.
    - Trigger event: fileCreated.
2. **Download File**
    - Retrieves the uploaded file from Google Drive.
    - Output: File binary.
3. **Default Data Loader**
    - Loads and parses the file into a format suitable for splitting and embedding.
    - Supports PDFs, text, and other document formats.
4. **Recursive Character Text Splitter**
    - Splits large text into manageable **chunks** (500 characters with 10% overlap).
    - Ensures embeddings capture semantic meaning without exceeding model limits.
5. **Google Gemini Embeddings**
    - Converts each text chunk into a **vector embedding**.
    - Embedding dimensions must match the Pinecone index configuration.
6. **Pinecone Vector Store**
    - Stores embeddings in Pinecone.Uses namespace (newfolder_n8n) for separation.

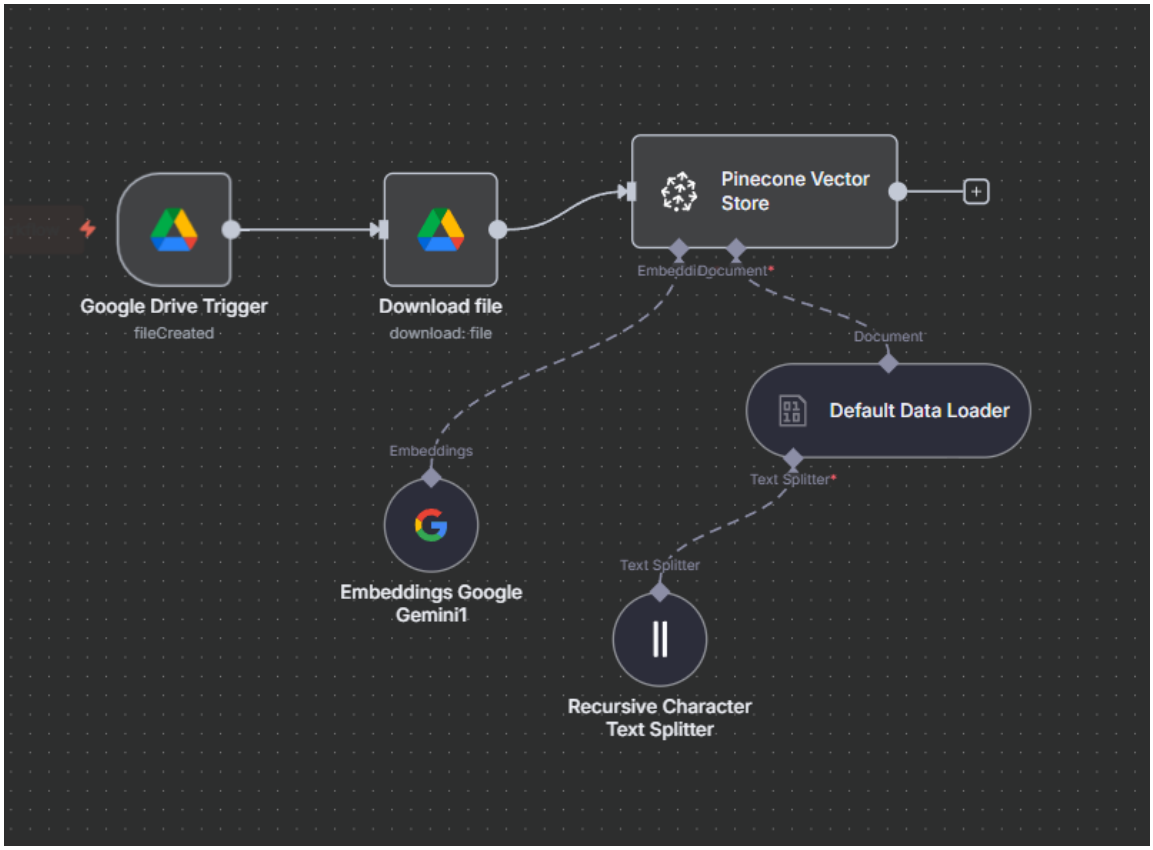

Fig1. Document Ingestion Pipeline

B. Retrieval & Chatbot Pipeline

1. **When Chat Message Received**
    - Captures incoming user query (chat Input).
2. **Google Gemini Chat Model**
    - Main LLM (chat completion).
    - Connected to AI Agent as the reasoning engine.
3. **AI Agent**
    - Uses the Gemini Chat Model to answer queries.
    - It use Pinecone vector store as tool to answer the queries.
4. **Pinecone Vector Store (Retriever Mode)**
    - Operation Mode: *Retrieve Documents (As Tool for AI Agent)*.
    - Uses the query embedding from Gemini to search Pinecone..

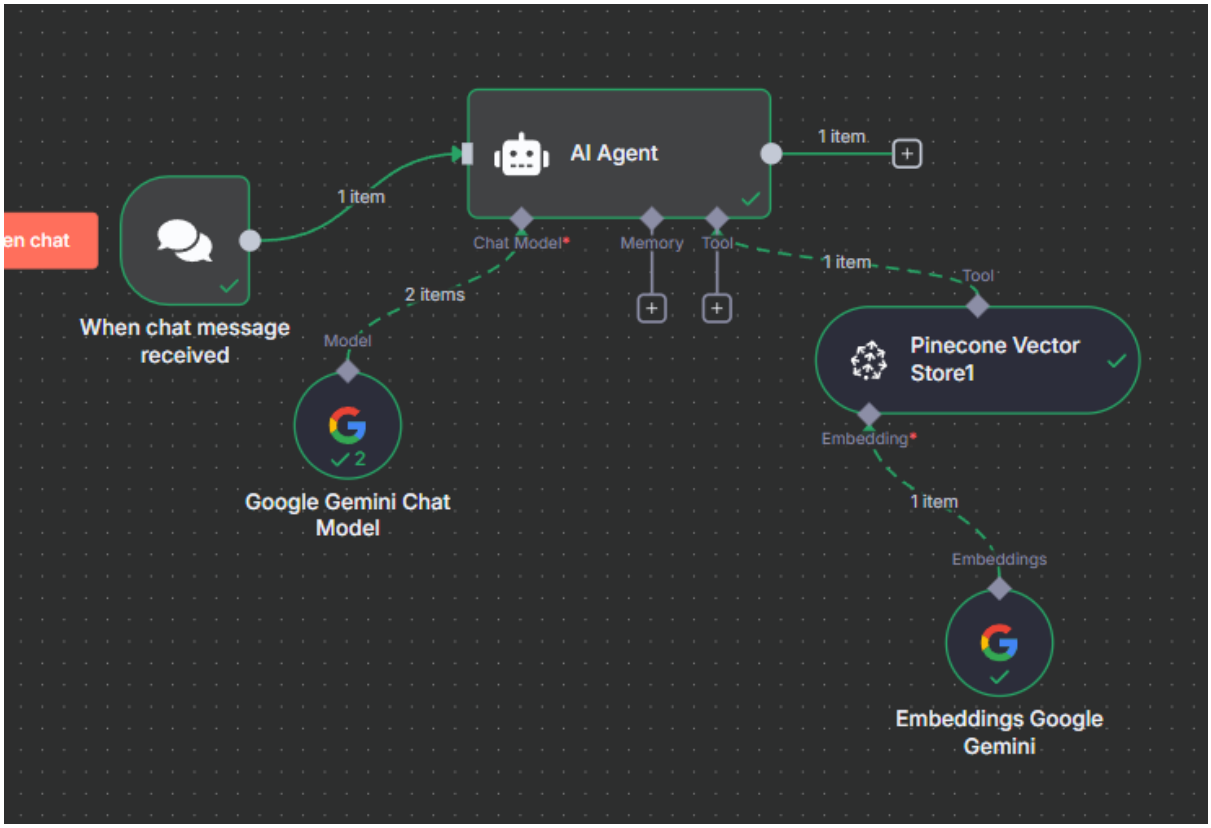

Fig 2 Retrieval & Chatbot Pipeline

3. Key Parameters

Document Chunking

- Chunk Size: 500 characters
- Overlap: 50 characters
- Purpose: Avoids loss of context between chunks.

Pinecone Settings

- Index: demo (example)
- Namespace: newfolder_n8n
- Limit (Top-K): Typically 4

AI Agent Settings

- Model: Google Gemini Chat
- Tools: Pinecone Vector Store (Retriever)
- Role: Respond to user queries with context-augmented answers.