Ingoude Company

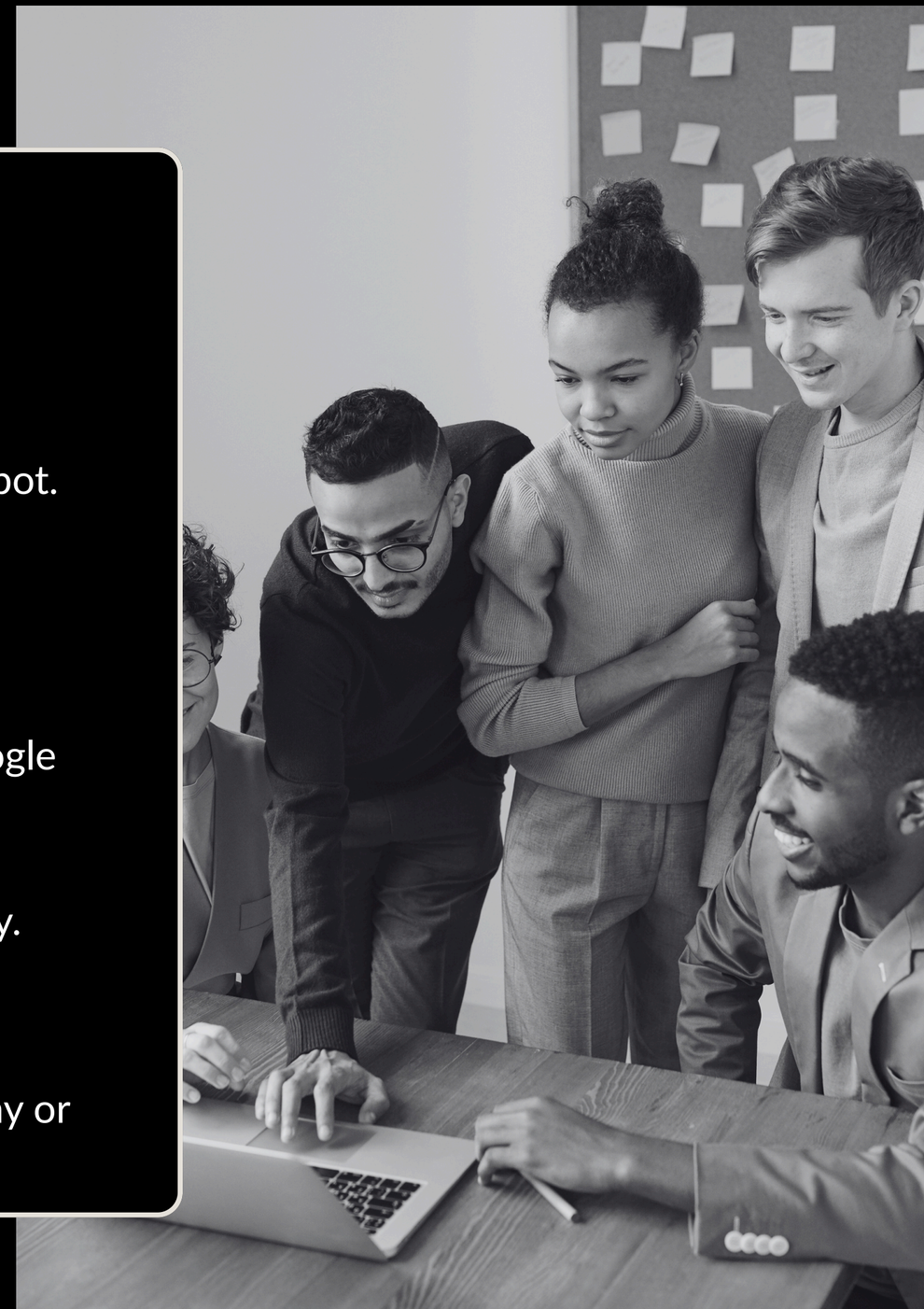
RAG Chatbot for Company Documents

Present by Smaran

TOPICS COVERED

INTRODUCTION

- The project implements a Retrieval-Augmented Generation (RAG)–based chatbot.
- Integrates Google Drive, Google Gemini, and Pinecone using n8n workflow automation.
- Enables users to upload, process, and summarize documents directly from Google Drive.
- Uses vector embeddings to store and retrieve document information efficiently.
- Provides AI-powered contextual answers and summaries for uploaded company or academic documents.



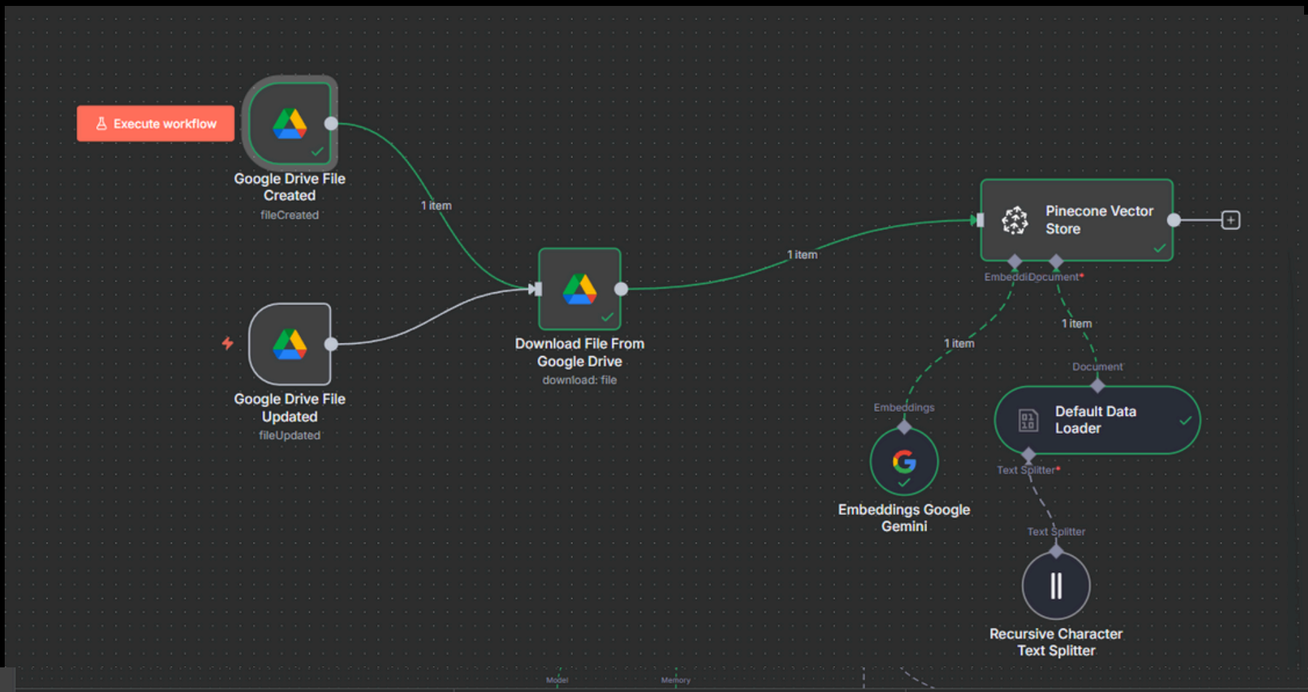
OBJECTIVE

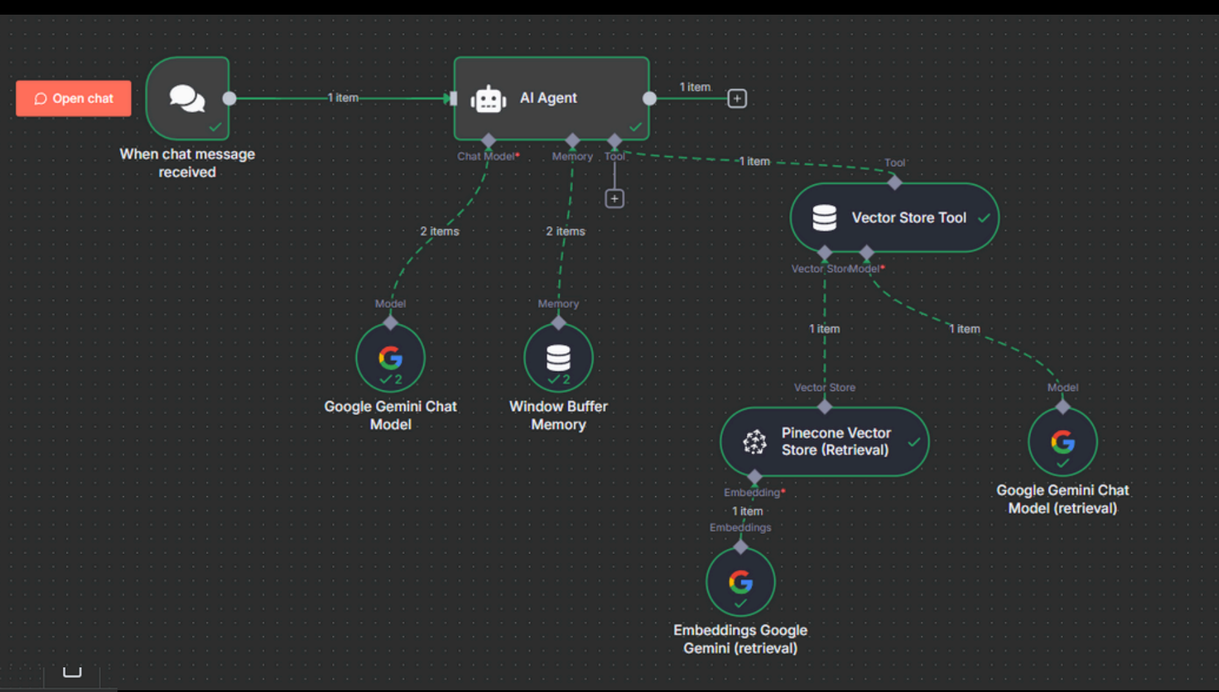
- To develop an AI-driven chatbot capable of reading and understanding company or academic documents.
- To integrate Google Drive for seamless document upload and access.
- To implement Retrieval-Augmented Generation (RAG) using Google Gemini for intelligent summarization and Q&A.
- To use Pinecone Vector Database for efficient storage and retrieval of document embeddings.
- To automate the document summarization and response generation process using n8n workflows.
- To ensure a no-code, scalable, and cloud-based solution for document analytics.
- To demonstrate the real-world application of Large Language Models (LLMs) in enterprise environments.

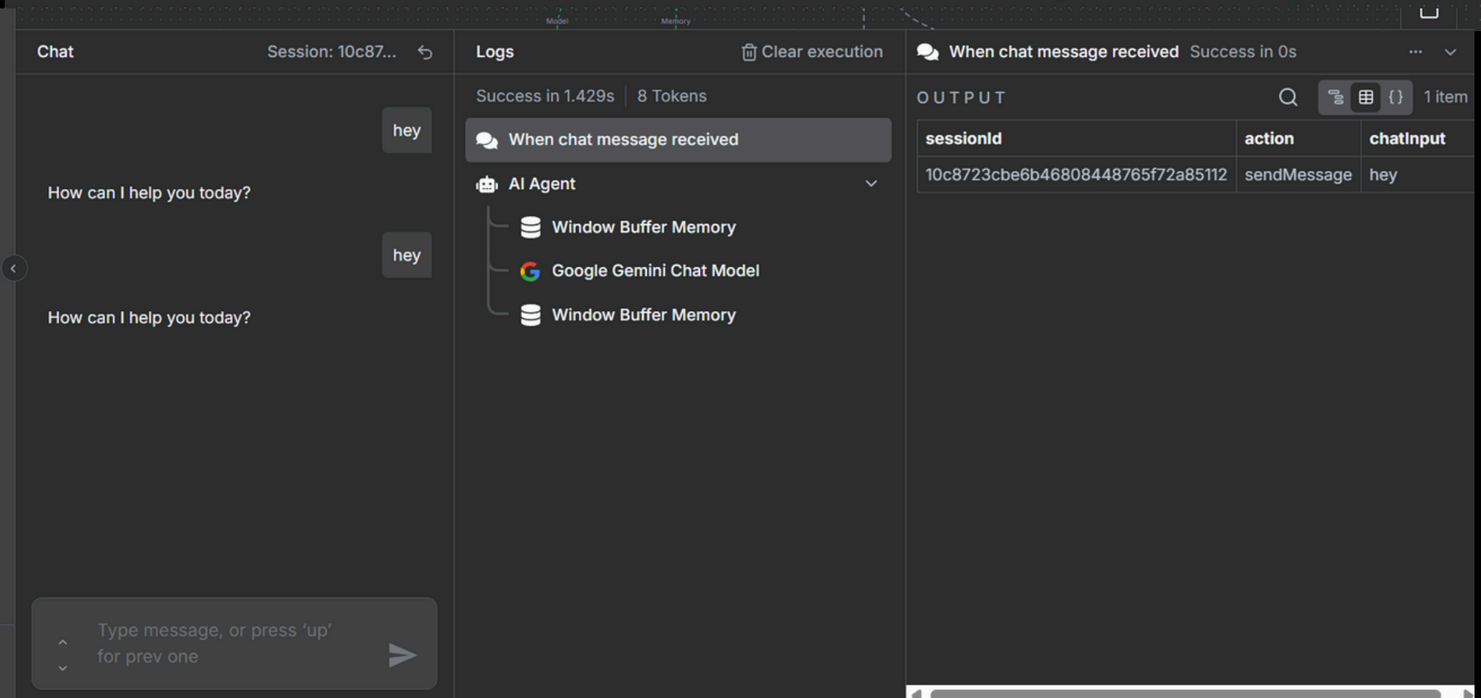
TECH STACK

Component	Technology / Tool Used	Purpose / Description
Workflow Engine	**n8n**	Automates the entire process through a no-code workflow builder
Language Model (LLM)	**Google Gemini 1.5 / Pro**	Generates summaries and answers based on retrieved content
Embeddings Model	**Gemini Embedding API (models/text-embedding-004)**	Converts text into numerical vectors for similarity search
Vector Database	**Pinecone**	Stores and retrieves document embeddings for RAG-based queries
File Storage	**Google Drive**	Stores uploaded PDF, DOCX, and TXT files
Data Loader	**Default Data Loader**	Extracts content from uploaded documents
Text Processor	**Recursive Character Text Splitter**	Breaks long documents into smaller chunks for embedding
Integration Layer	**API Connections (n8n Nodes)**	Enables seamless communication between Gemini, Drive, and Pinecone
AI Agent	**Gemini AI Agent Node**	Handles query input, retrieval, and response generation
Output Manager	**Google Drive – Create File Node**	Automatically saves generated summaries back to Drive

SYTEM ARCHITECTURE







DETAILED COMPONENT EXPLAINATION

Workflow Node	Detailed Explanation
Google Drive – File Uploaded / Created	Detects when a new document is uploaded to the connected Google Drive folder. This serves as the **trigger point** for the entire workflow. It passes the file ID and metadata to the next node.
PDF Extract / Default Data Loader	Takes the uploaded PDF or DOCX file and **extracts the text content** into a readable format. This step ensures that non-text elements (like tables or images) are ignored, focusing only on textual information.
Recursive Character Text Splitter	**Splits the extracted text** into smaller overlapping chunks (e.g., 1000 characters with 200 overlap). This helps preserve meaning and context during vectorization.
Gemini Embeddings Node	Converts the text chunks into **semantic vector embeddings** using Google's models/text-embedding-004. These embeddings capture the conceptual meaning of the text.
Pinecone Vector Store (Insert)	Stores all embeddings into Pinecone's vector database. Enables **fast semantic search** for relevant text when a user asks a question or requests a summary.
Vector Store Tool (Retriever)	Acts as the **bridge between Pinecone and the AI Agent**. When a user query arrives, it retrieves the most relevant stored embeddings (document chunks) from Pinecone.
AI Agent (Google Gemini)	The **central decision-maker**. Takes the retrieved context from Pinecone, interprets the user's prompt, and generates an accurate response or summary using Gemini's reasoning.
Google Drive – Create File Node	Takes the final AI response and **creates a new .txt file** in the connected Google Drive. The file is automatically named (often with a timestamp) and stores the chatbot's answer or summary.

SYTEM ARCHITECTURE

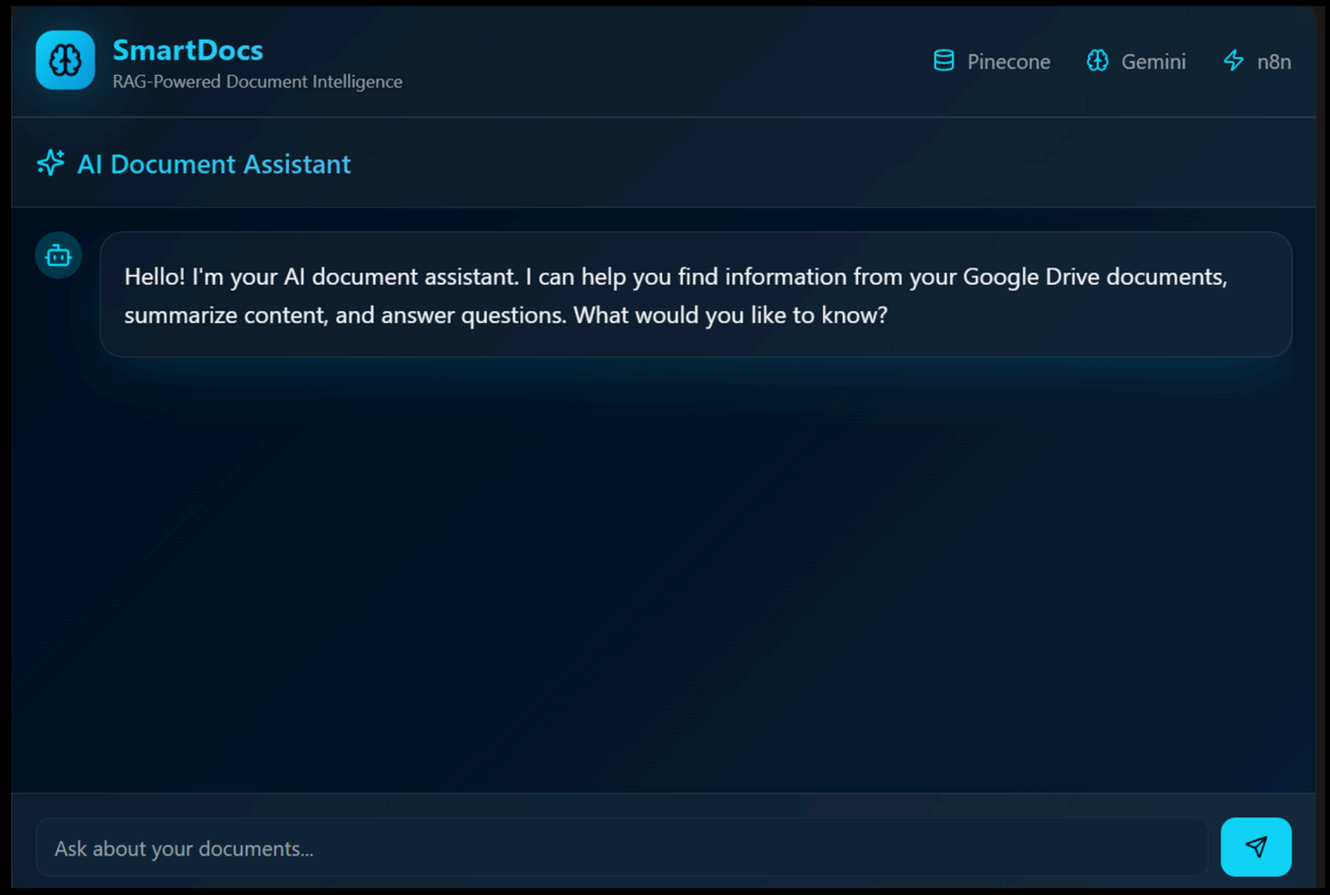
Workflow	Detailed Explanation
When Chat Message Received (Trigger)	Acts as the **starting point for user interaction**. When a user types a question or command (e.g., *“Summarize DBMS Unit 1”*), this node triggers the retrieval and response flow.
Google Gemini Chat Model	The **language model** that interprets user messages, understands context, and communicates with other nodes to generate intelligent responses. It uses the Gemini API to process inputs dynamically.
Vector Store Tool (Retriever)	Acts as a **connector between Pinecone and the AI Agent**. When the AI Agent receives a user question, this tool fetches the most relevant document embeddings from Pinecone based on semantic similarity.
Pinecone Vector Store (Retrieval)	Performs the **vector similarity search** operation. It retrieves document chunks from Pinecone that are most similar to the user’s question embedding. These chunks act as the “context” for Gemini’s answer.
AI Agent (Google Gemini)	The **core intelligent node**. Combines the retrieved document context (from Pinecone) with the user’s query and generates an accurate, context-aware answer or summary using Gemini’s large language model capabilities.
Window Buffer Memory (Optional)	Maintains **short-term conversational memory**, allowing the chatbot to remember previous interactions in a single session for more natural, flowing dialogue.
Google Drive – Create File from Text	Takes the AI Agent’s generated response and **creates a new text file** in Google Drive. The file typically contains the summarized or queried output, saved automatically with a timestamped name (e.g., Response_2025-11-01.txt).

AI AGENT

Aspect	Detailed Explanation
Purpose	The **AI Agent** is the central intelligence of the workflow. It connects all components — receiving user input, retrieving relevant data from Pinecone, and generating meaningful, context-aware answers using **Google Gemini**.
Core Function	Implements **Retrieval-Augmented Generation (RAG)** — combines retrieved document chunks from the **Vector Store Tool** with Gemini's reasoning to produce summaries or answers grounded in uploaded data.
Model Used	**Google Gemini 1.5 Pro / Flash**, depending on workflow configuration. Capable of understanding natural language and generating detailed responses.
Input Source	Accepts input from the **Chat Trigger Node** (user message) and contextual data from the **Vector Store Tool** (retrieved document vectors).
Prompt Template	Uses two parts:• **System Prompt:** Defines the agent's role (e.g., "You are a helpful assistant that summarizes company documents").• **User Prompt:** Dynamic message from user (e.g., "Summarize DBMS Unit 1").
Tool Integration	Connected to the **Vector Store Tool** for contextual search and optionally to **Window Buffer Memory** for maintaining conversation history.
Output	Generates summarized text or Q&A responses and passes it to the **Create File Node** to be saved in Google Drive.
Error Handling	If Pinecone retrieval fails, the AI Agent falls back on Gemini's model knowledge to generate a general answer.

FUTURE SCOPE

- Integrate more data sources like Notion, Google Sheets, and Gmail for broader document access.
- Implement chat memory to enable continuous, context-aware conversations.
- Develop a dashboard interface for monitoring summaries and AI responses.
- Enhance accuracy and scalability by fine-tuning Gemini and optimizing Pinecone retrieval.



CONCLUSION

- Developed an AI-powered RAG Chatbot that integrates Google Drive, Gemini, and Pinecone.
- Automated the process of document reading, summarization, and contextual answering.
- Implemented a vector-based retrieval system for efficient and accurate information access.
- Achieved a fully cloud-based, no-code workflow using n8n for seamless automation.
- Enhanced productivity and data accessibility through AI-driven document intelligence.
- Demonstrated the real-world application of LLMs in enterprise and academic knowledge management.
- Established a foundation for future expansion, allowing integration with more data sources.
- Showcased how automation and AI together can transform document analytics and knowledge retrieval.
- Provided a scalable and practical solution for intelligent document processing in modern organizations.

THANK YOU