

## Lab8\_2: RAG Chatbot

**Retrieval-Augmented Generation (RAG)** is a way to build a chatbot that answers questions using your own documents. Instead of relying only on the language model, the workflow first **retrieves** the most relevant text from a document collection, then **generates** an answer based on that retrieved content.

### Document

For this lab, the RAG chatbot will use a short Thai educational brochure titled “**ความดันโลหิตสูง... อันตรายกว่าที่คุณคิด**” produced by the **สำนักงาน Mayor กรุงเทพมหานคร** (Bangkok Metropolitan Administration).

The brochure is **2 pages** and is structured as practical guidance for the public, covering: **What hypertension is** and the diagnostic threshold ( $\geq 140/90$  mmHg) plus why uncontrolled blood pressure increases risk of complications.

- **Blood pressure categories table** (e.g., normal / elevated / stage levels) with suggested actions such as lifestyle control, consulting a doctor, or urgent medical attention for severe levels.
- **Key contributing factors** (e.g., age, stress/emotions, heredity, salty diet, urban lifestyle) and common **symptoms** (often none, but may include headache, dizziness, fatigue).
- **Complications** described as short-term (urgent/severe symptoms) and long-term organ damage (often after many years).
- **Emergency warning signs** and recommended immediate actions (e.g., what to do if BP is abnormally high/low, when to seek urgent care).
- **Prevention and control checklist**, including target BP guidance by age group, reducing salty foods, exercise, stress reduction, medication adherence, and annual screening for complications.

In the RAG chatbot, this brochure will function as the “ground truth” reference: the bot should retrieve relevant sections (tables, warning signs, prevention steps) and answer questions **based on the brochure content**.

## Expected Behaviour

Prompt	Response
Based on the provided brochure, what counts as hypertension (include the BP threshold), and what are the key first steps to prevent or control it? Please cite the relevant parts of the document.	<p>Key points:</p> <ul style="list-style-type: none"> <li>● <b>Hypertension Definition:</b> Hypertension is diagnosed when blood pressure consistently measures above 140/90 mmHg, as defined by the World Health Organization.</li> <li>● <b>Blood Pressure Targets:</b> <ul style="list-style-type: none"> <li>○ For individuals under 50 years old, blood pressure should ideally be below 130/80 mmHg.</li> <li>○ For those 50 years and older with chronic conditions, the target is below 140/90 mmHg.</li> <li>○ For individuals 60 years and older, blood pressure should be below 150/90 mmHg.</li> </ul> </li> <li>● <b>Key First Steps for Prevention and Control:</b> <ul style="list-style-type: none"> <li>○ Manage blood pressure according to age-specific targets.</li> <li>○ Reduce sodium intake, including salt, MSG, and seasonings, and avoid pickled foods.</li> <li>○ Eat appropriate portion sizes, avoid heavy meals before sleep, and limit frequent large gatherings or buffets.</li> <li>○ Cease smoking and all alcohol consumption.</li> <li>○ Engage in physical activity whenever possible, such as 30 minutes of continuous housework.</li> <li>○ Manage stress, maintain a positive outlook, practice meditation, watch movies, listen to music, or play sports.</li> <li>○ Take prescribed medications consistently without adjusting doses or stopping independently.</li> <li>○ Undergo annual screenings for complications.</li> </ul> </li> </ul> <p>Evidence:</p> <ul style="list-style-type: none"> <li>● "โรคความดันโลหิตสูง (Hypertension) ... โดยองค์การอนามัยโลกกำหนดให้ผู้ที่มีความดันโลหิต มากกว่า 140/90 มม.ปรอท ถือว่าเป็นโรคความดันโลหิตสูง"</li> <li>● "1. ควรควบคุมความดันโลหิต ... อายุต่ำกว่า 50 ปี ควรต่ำกว่า 130/80 มม.ปรอท ... อายุ 50 ปีขึ้นไป และมีโรคเรื้อรัง ความดันโลหิตต้อง ต่ำกว่า 140/90 มม.ปรอท ... อายุ 60 ปีขึ้นไป ต่ำกว่า 150/90 มม.ปรอท"</li> <li>● "2. ลดอาหารเค็มในอาหาร โดยค้่นึงถึงปริมาณเกลือ ผงชูรส เครื่องปรุงรสต่างๆ เลี่ยงของหมักดอง"</li> <li>● "3. กินอาหารในปริมาณที่พออิ่ม งดอาหารหนักก่อนนอน และไม่ควรไปงานเลี้ยงหรือกินบุฟเฟ่ต์บ่อย"</li> <li>● "4. งดสูบบุหรี่ และงดดื่มเครื่องดื่มแอลกอฮอล์ทุกชนิด"</li> <li>● "5. ออกกำลังกายทุกครั้งที่มีโอกาส หรือ ที่ทำงานบ้านต่อเนื่อง ประมาณ 30 นาที"</li> <li>● "6. หลีกเลี่ยงความเครียด ท้าจิตใจให้แจ่มใส ผักผลไม้สด ดูนั่ง ฟังเพลง หรือเล่นกีฬา"</li> <li>● "7. กินยาตามแพทย์สั่งอย่างสม่ำเสมอ ไม่ปรับขนาดยาหรือหยุดยาเอง"</li> <li>● "8. ตรวจคัดกรองภาวะแทรกซ้อนทุกปี"</li> </ul>
According to the brochure, what warning signs suggest a hypertensive emergency, and what should someone do immediately? Answer in clear steps and cite the document sections used.	<p>Key points:</p> <p>Warning signs of a hypertensive emergency, which occur in individuals with severe high blood pressure, include:</p> <ul style="list-style-type: none"> <li>● Severe headache, particularly at the back of the head.</li> <li>● Blurred vision due to swelling of the optic nerve.</li> <li>● Drowsiness progressing to seizures.</li> <li>● Severe chest pain in the center, potentially radiating to the neck or left arm.</li> <li>● Easy fatigue and inability to lie flat.</li> <li>● Sudden, severe pain in the chest or abdomen that radiates to the back.</li> <li>● Sudden numbness or weakness in the arms and legs.</li> <li>● Frequent dizziness upon standing.</li> <li>● Abnormal heart palpitations with an irregular pulse.</li> <li>● Unusual fatigue.</li> <li>● Waking up at night with shortness of breath and inability to lie flat.</li> <li>● Unusual paleness accompanied by swollen feet in the afternoon and frequent nighttime urination.</li> <li>● Headaches and blurred vision not due to cataracts.</li> </ul>

	<ul style="list-style-type: none"> <li>• Confusion, restlessness, or stupor leading to seizures.</li> <li>• Numbness or weakness on one side of the body, or temporary blindness in one eye that resolves within 24 hours.</li> </ul> <p>Immediate actions:</p> <ol style="list-style-type: none"> <li>1. <b>Seek Medical Attention Immediately:</b> After initial self-care, quickly go to the nearest hospital or call an emergency ambulance.</li> <li>2. <b>Measure Blood Pressure (if possible):</b> If you have a home blood pressure monitor, check if your blood pressure is abnormally high or low.</li> <li>3. <b>For High Blood Pressure (due to stopping medication):</b> Take your usual prescribed medication.</li> <li>4. <b>For Shortness of Breath and Inability to Lie Flat:</b> Sit or lie with your head elevated.</li> <li>5. <b>For Low Blood Pressure (possibly due to diarrhea, vomiting, gastrointestinal bleeding, or overdose of medication):</b> Lie flat, drink electrolyte solution as much as you can, and temporarily stop taking blood pressure medication.</li> </ol> <p>Evidence:</p> <ul style="list-style-type: none"> <li>• "อาการของผู้ที่มีความดันโลหิตสูง ... ภาวะแทรกซ้อนจากโรคความดันโลหิตสูง ... ระยะสั้น เกิดในผู้ที่มีความดันโลหิตสูง ระดับรุนแรงมีอาการ ปวดศีรษะบริเวณท้ายทอย ... ตามัวจากข้อประสาทตาบวม ... ซึมลงและชัก ... เจ็บหน้าอก กลางหน้าอก ร้าวมาที่คอหรือแขนซ้าย ... เหนื่อยง่ายนอนราบไม่ได้ ... ปวดบริเวณหน้าอกหรือท้องทะลุหลังอย่างเฉียบพลัน ... แขนขาชาหรืออ่อนแรงเฉียบพลัน"</li> <li>• "2 อาการที่มีความผิดปกติและต้องรีบแก้ไข ... เหนื่อยง่ายผิดปกติ ... ตื่นกลางดึก ลูกขึ้นมากอหอบ นอนราบไม่ได้ ... ซีดผิดปกติร่วมกับเท้าบวมตอนบ่าย ปัสสาวะบ่อยกลางคืน ... ปวดศีรษะ ตามัวโดยไม่ได้เป็นต่อกระจก ... สับสนกระวนกระวาย ซึมจนถึงชัก ... แขนขา ชา อ่อนแรงซีกเดียวกัน หรือมองไม่เห็นชั่วคราวข้างหนึ่ง และหายได้เองใน 24 ชม. ... เจ็บแน่นหน้าอกตรงกลาง อาจร้าวมาที่คอหรือแขนซ้าย ... ปวดบริเวณหน้าอกหรือท้อง อย่างรุนแรงเฉียบพลัน ร้าวไปข้างหลัง ... ลูกขึ้นหน้ามืดบ่อยๆ ... ใจสั่นผิดปกติ ชีพจรเต้นไม่เป็นจังหวะ"</li> <li>• "หลังดูแลตนเองเบื้องต้น ให้รีบไปพบแพทย์ ที่สถานพยาบาลใกล้บ้าน หรือโทรเรียกรถฉุกเฉิน"</li> <li>• "หากมีเครื่องวัดความดันที่บ้าน ให้วัดว่าสูงหรือต่ำผิดปกติหรือไม่"</li> <li>• "กรณีความดันโลหิตสูง เพราะหยุดยาไป ให้กินยาเดิม"</li> <li>• "กรณีเหนื่อย นอนราบไม่ได้ ให้นอนหัวสูง"</li> <li>• "กรณีความดันโลหิตต่ำ อาจมีสาเหตุจาก ท้องเสีย อาเจียน เลือดออกทางเดินอาหาร หรือได้รับยาลดความดันโลหิตเกินขนาด ให้นอนราบ และดื่มเกลือแร่ปริมาณเท่าที่ดื่มได้ และ หยุดยาลดความดันโลหิตไปก่อน"</li> </ul>
--	--

## Prerequisites

Google Gemini API Key

### A) Create a Google Gemini API key (Google AI Studio)

#### What you need

- A Google account
- Access to Google AI Studio (you may need to accept Terms of Service on first use)

#### Steps

1. Open the **Google AI Studio API Keys** page and sign in.
2. Click **Create API key** (or **Create or view a Gemini API Key**).
3. Copy the key and store it securely (do not share it or paste it into chat messages).

#### Where you will use it in this lab

- In n8n, create a credential for **Google Gemini API account** and paste the key into the credential (so nodes can call the model securely).

Google AI Studio – API Keys: <https://aistudio.google.com/app/apikey>

Gemini API keys guide (official): <https://ai.google.dev/gemini-api/docs/api-key>

## B) Activate n8n Cloud free trial (managed n8n)

### What you need

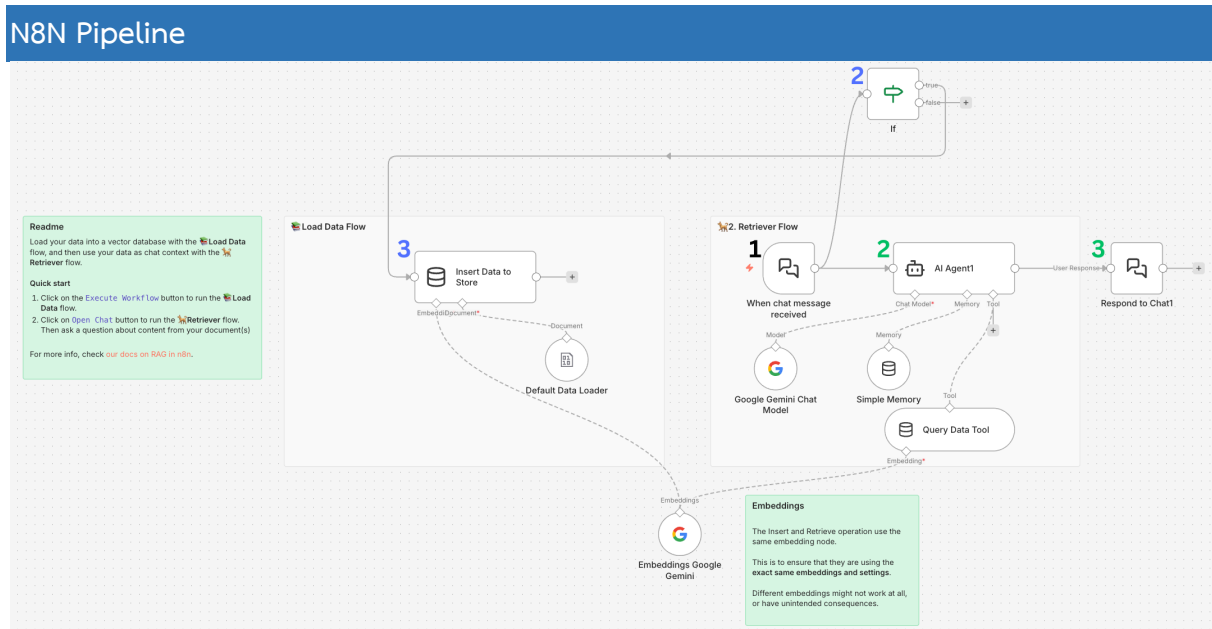
- An email address for registration

### Steps

1. Go to the n8n Cloud registration page and create a new workspace.
2. Your Cloud free trial includes **14 days** of Pro-plan features, with a limit of **1000 executions** (and Starter-level compute).

n8n Cloud trial registration: <https://app.n8n.cloud/register>

n8n Cloud free trial docs (official): <https://docs.n8n.io/manage-cloud/cloud-free-trial/>



### 1) Chat Trigger Node — When chat message received (Node 1)

This node is the **entry point** of the RAG chatbot. It starts the workflow whenever a user sends a message in the n8n chat UI. In this lab, we use it to support two actions:

**Upload documents** (to build/update the RAG knowledge base), and

**Ask questions** (to retrieve relevant passages and generate an answer).

### Configurations:

- Make Chat Publicly Available: On
- Mode: Hosted Chat

- Options → Allow File Uploads: On
- Options → Response Mode: Using Response Nodes

After the chat trigger, the pipeline routes into two parts:

1. **Load Data Flow (blue)**: used when the user uploads document(s) to be embedded and stored in the vector database.
2. **Retriever Flow (green)**: used when the user asks a question; the agent retrieves relevant chunks and responds via **Respond to Chat**.

---

### Load Data Flow (blue)

#### 2) If Node — Route: Upload (Load Data Flow)

This **If** node is used as a **router** right after the chat trigger. It checks whether the incoming chat message includes an **uploaded file**.

**True path → Load Data Flow (blue)**: a file was uploaded, so we load/embed the document and insert it into the vector store.

**False path → Retriever Flow (green)**: no file upload, so we treat the message as a normal question and run RAG retrieval + answering.

#### Configurations

- **Conditions**: The node uses **OR** logic (either condition can trigger the True branch).
  - **Condition 1 (detect any binary attachment)** Expression:

```
{{ Object.keys($binary || {}).length > 0 }}
```

- **Condition 2 (detect a specific binary field)** Expression:

```
{{ !!$binary?.data }}
```

#### 3) Insert Data to Store

This node takes the **uploaded file(s)** from the chat trigger, converts them into **document chunks**, creates **embeddings**, and then **stores** them in the vector store. This is what “loads” your document into the RAG knowledge base so it can be retrieved later.

#### Configurations

- Operation Mode: Insert Documents
- Memory Key: From list → `vector_store_key`
- Embedding Batch Size: 200
- Clear Store: On

## Required attachments

1. Document \* → Default Data Loader
2. Embedding \* → Embeddings Google Gemini

### Attachment A: Default Data Loader (Document input)

Loads the uploaded file from the chat message (binary file upload), extracts text, and **splits** it into chunks suitable for embedding.

#### Configurations

- Type of Data: Binary
- Mode: Load All Input Data
- Data Format: Automatically Detect by Mime Type
- Text Splitting: Simple

### Attachment B: Embeddings Google Gemini (Embedding input)

Generates vector embeddings for each document chunk so the vector store can perform semantic similarity search during retrieval.

#### Configuration

- Credential to connect with: Google Gemini API account
- Model: models/gemini-embedding-001

---

## Retriever Flow (green)

### 2) AI Agent

AI Agent is the “brain” of the chatbot. It receives the user’s chat message, decides when to call the retrieval tool, and then writes the final answer based on the retrieved document text.

#### Configurations

- Source for Prompt: Connected Chat Trigger Node
- Prompt

`{{ $json.chatInput }}`

- Require Specific Output Format: Off

**Options → System Message:**

Role: Medical education assistant for students.

You have access to a “Query Data Tool” that retrieves text from the uploaded document.

Rules:

For any question that might be answered from the uploaded document, you MUST call Query Data Tool before answering.

Base your answer only on retrieved text. If retrieval returns nothing relevant, say the document doesn’t cover it.

Output format:

Key points (bullets)

Evidence (2–4 short snippets from retrieved text)

Safety:

No diagnosis, prescribing, or patient-specific treatment plans.

If the user asks for personal medical advice, recommend consulting a clinician.

- Max Iterations: 2

#### Attachments

- a Chat Model (Gemini Chat Model)
- Memory (Simple Memory)
- Tool (Query Data Tool)

#### Attachment A: Query Data Tool (Vector Store Retriever Tool)

This node is the retrieval component. It searches the vector store using embeddings to return the most relevant document chunks. The AI Agent calls this tool automatically when it needs evidence from the document.

#### Configurations

- Operation Mode: Retrieve Documents (As Tool for AI Agent)
- Name: knowledge\_base
- Memory Key: From list → vector\_store\_key
- Limit: 4
- Include Metadata: On

#### Attachment A.1 Embedding: (Gemini embedding node)

The retriever must use the same embedding model as the indexing step to ensure similarity search works correctly.

### 3) Respond to Chat

This node sends the **final answer** back to the n8n chat UI. It is required whenever the Chat Trigger node's **Response Mode** is set to **Using Response Nodes**.