SkillNavora · Written by SkillNavora · Plain-English version

RAG: How an AI Finds the Right Facts

This guide explains one idea: how an AI can answer questions from your own documents. You do not need any background.

> **The big idea:** before the AI answers, a search finds the right paragraph in your documents. The AI reads that paragraph and then answers.

1. Why does an AI need help?

An AI learned from a large collection of text. It has three weak spots:

Problem	Simple example
It does not know your private files	"How many leave days do I get?" It guesses.
It stops learning on a date	"What changed in last week's policy?" It cannot know.
It can make things up	It gives a confident answer that is wrong.

2. Three ways to help the AI

Way	What it means	Good for
Fine-tuning	Training the AI again on your examples. This changes what it knows.	Style and format, like always writing in a set way
In-context learning	Showing a few examples in your question. Nothing is trained.	Teaching a pattern quickly
RAG	Searching your documents first, then giving the best paragraphs to the AI.	Facts that change, with a source to check

> **Easy rule:** for facts, use RAG. For tone, use clear instructions. Use fine-tuning only when the other two do not work.

3. The four steps of RAG

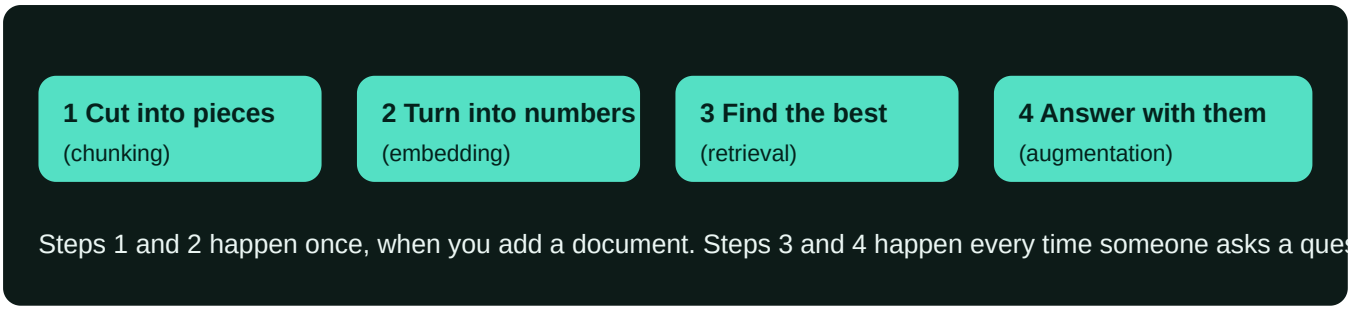


Figure 1. The four steps, in order.

Step 1: cut the document into pieces

A long document is cut into small pieces called **chunks**. A chunk is small enough to search, and big enough to make sense.

Example from the class: a 519-letter paragraph was cut into chunks of about 120 letters, and each chunk shares about 15 letters with the next one. This is an example, not a rule. Start with a size that fits your document, then test it.

Why the overlap? A sentence cut in half can still appear whole in a neighbouring chunk.

Step 2: turn each piece into numbers

An **embedding** is a list of numbers that stands for the meaning of a piece of text. Pieces with a similar meaning get similar numbers.

Use **the same embedding tool** for your pieces and for the questions. Otherwise the numbers cannot be compared.

The numbers are kept in a **vector store**. This is a database that holds each piece, its numbers and its source name.

Step 3: find the best pieces

When someone asks a question, the question also becomes numbers. The search finds the pieces with the closest numbers. This is called **similarity**.

Example: the question "How does our leave policy work?" is closest to "Annual leave rules" (score 0.91) and "Sick leave policy" (score 0.88). Those two pieces are kept. The "Office dress code" piece scores much lower, so it is left out.

The number of pieces kept is called **top K**. If K is 2, the two best pieces are kept.

Step 4: give the pieces to the AI

The AI gets three things, in this order:

1. **The rule:** "Answer only from the pieces below. If they do not have the answer, say I do not know."
2. **The pieces,** with their source names.
3. **The question.**

```
Rule: Answer only from the context. If it is missing, say "I don't know".

Context:
[Source: Annual leave rules, page 2] Employees get 18 days of leave a year.
[Source: Sick leave policy, page 1] Tell your manager on the same day.

Question: How does our leave policy work?
```

Notice the rule comes first. The AI reads the rule before the facts.

4. Check your answers

Make a list of 20 questions. For each one, write the answer you expect and the page it comes from. Then check three things:

1. **Did the search find the right page?**
2. **Does the answer only use facts from that page?**
3. **Does the AI say "I don't know" when the page does not have the answer?**

If the search is wrong, change the piece size or the number K. If the answer adds facts, make the rule stricter.

5. Questions and answers

Q1. A formal Hindi chatbot. RAG or fine-tuning?
Start with clear instructions for the tone. Use RAG for facts. Fine-tune only if the tone still fails.

Q2. Can you use RAG and fine-tuning together?
Yes. Fine-tuning sets the style. RAG brings the facts.

Q3. RAG for news from the last hour?
Only if new articles are added to the search within minutes. Fine-tuning is too slow for this.

Q4. RAG over 500 PDFs gives partly right answers. What is wrong?
Usually the search. Check whether the right piece is in the top K. Check the piece size, the number K, and whether the PDFs are readable text, not scanned pictures.

Q5. RAG works in English but fails in Hindi. Why?
The embedding tool may not understand Hindi well. Use one that handles many languages, and test with Hindi questions.

6. Word bank

Word	Meaning
RAG	Search your documents first, then answer.
Chunk	A small piece of a document.
Embedding	Numbers that stand for meaning.
Vector store	A database that holds pieces and their numbers.
Top K	How many best pieces the AI gets.
Hallucination	A confident answer that is wrong.

7. Check yourself

1. Which step happens once, when you add a document?
2. Why must the same embedding tool be used for pieces and questions?
3. What does top K = 2 mean?
4. What should the AI say when the pages do not have the answer?

Answers: 1. Steps 1 and 2 (cutting and turning into numbers). 2. Otherwise the numbers cannot be compared. 3. The two best pieces are given to the AI. 4. "I don't know."