A 20-MINUTE BEGINNER GUIDE

01 / RETRIEVE

02 / AUGMENT

03 / GENERATE

# RAG: Giving LLMs the Right Context

How a company handbook assistant finds evidence before it answers.

`question` → **`evidence`** → `answer`

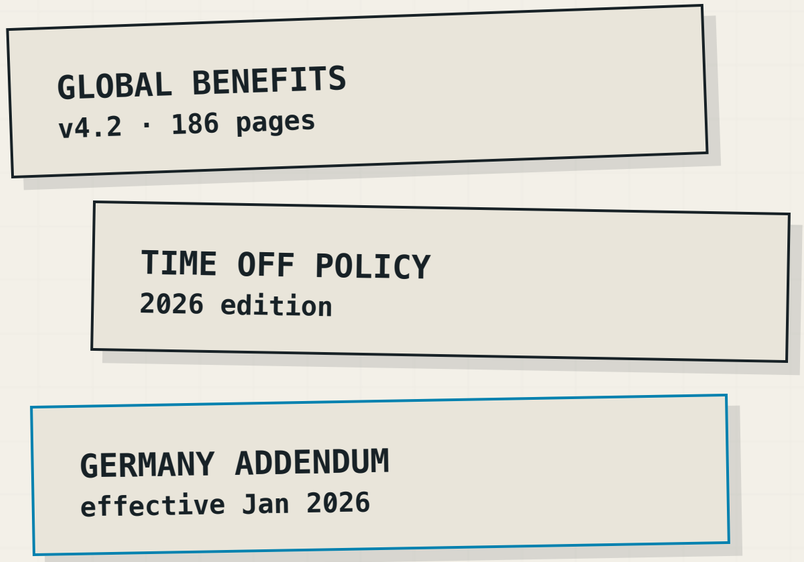
THE RUNNING EXAMPLE

# One ordinary employee question

EMPLOYEE · BERLIN OFFICE

“Can I carry over unused vacation days into next year?”

channel: handbook assistant • needed: policy + location



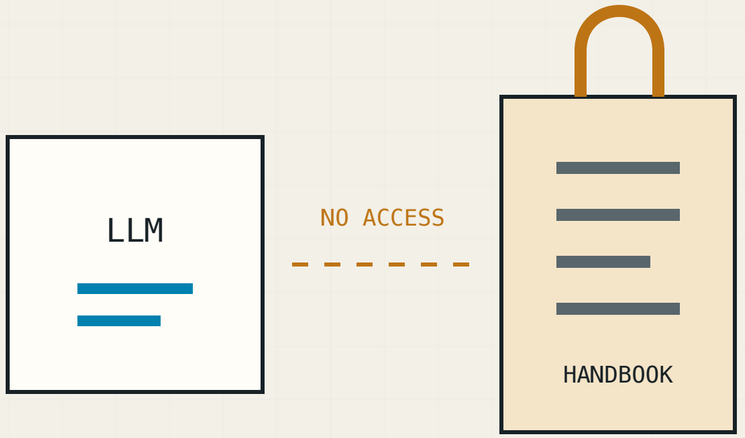
THE KNOWLEDGE GAP

# The model cannot know what it was never given

01 Your handbook is **private**, not part of public training data.

02 Policies **change** after a model's training cutoff.

03 A fluent model may still **guess** when evidence is missing.



THE CORE MOVE

# Retrieve first. Generate second.

RAG adds an evidence-gathering step before the LLM writes.

INPUT / 01

## Employee question

What policy applies to me?

→

RETRIEVAL / 02

## Find relevant text

Search the handbook for likely evidence.

→

GENERATION / 03

## Answer with context

Use the evidence; cite the source.

END-TO-END

# The full RAG pipeline



### 1. Question

Capture the user's need and metadata.



### 2. Retrieve

Rank handbook chunks by relevance.



### 3. Augment

Place selected evidence in the prompt.



### 4. Generate

Ask the LLM to answer from context.



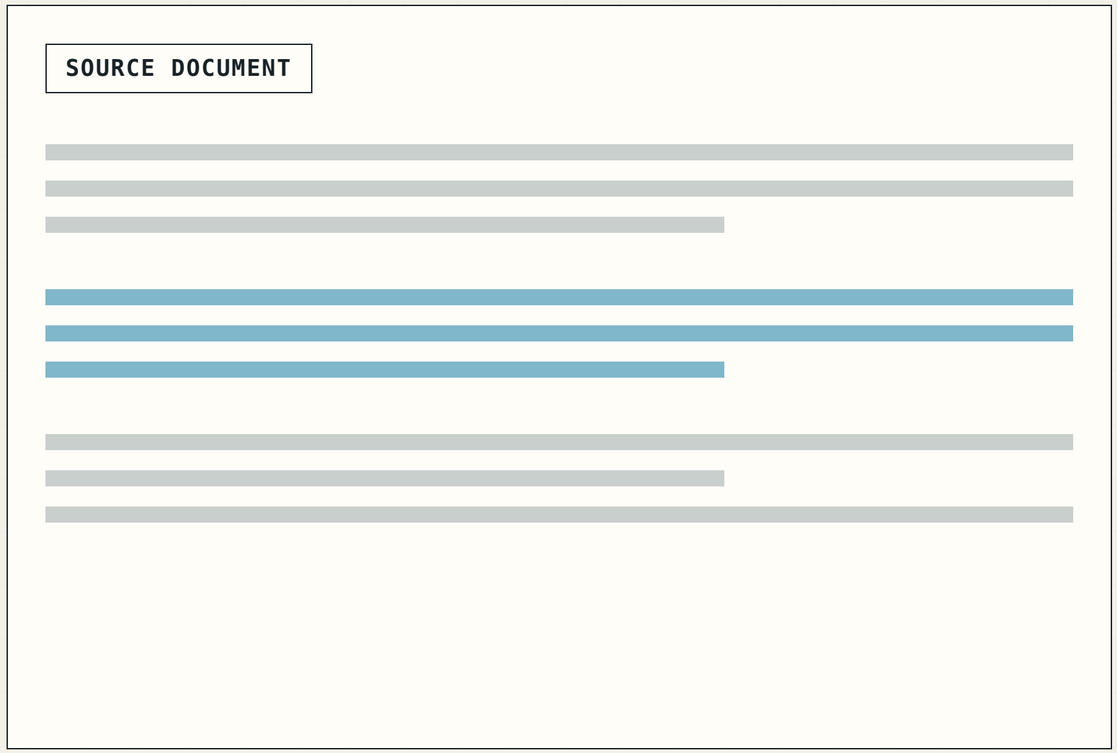
### 5. Cite

Return the answer with traceable sources.

RAG is a system around the model — not a different kind of model.

BEFORE ANYONE ASKS

# Prepare the handbook in useful pieces




**CHUNK 014**

Vacation allowance and annual reset.

**CHUNK 015**

Carryover rules and local exceptions.

**CHUNK 016**

Manager approval and deadlines.

**CHUNK 017**

Leaving the company and payout.

Keep each chunk coherent; preserve its title, section, version, and location metadata.

MEANING, MADE SEARCHABLE

# Embeddings turn text into coordinates

Text with similar meaning lands near other similar text — even when the exact words differ.

CHUNK 015

“Unused leave may be carried into Q1…”

`[ 0.12, −0.48, 0.81, … ]`

**“carry over vacation”**
close in meaning

**“unused leave next year”**
also close

**“laptop reimbursement”**
far away

AT QUESTION TIME

# Retrieve the top-k closest chunks

QUERY EMBEDDING

Can I carry over unused vacation?

```
[ 0.09, −0.43,
0.77, … ]
```

```
k = 3
filter: country = DE
```

| | | |
|---|---|---|
| **#1** | Germany addendum · carryover exception | **0.92** |
| **#2** | Time off · unused leave policy | **0.87** |
| **#3** | Time off · manager approval deadline | **0.78** |
| **#4** | Benefits · sabbatical eligibility | **0.41** |
| **#5** | Expenses · home office equipment | **0.13** |

“Top-k” simply means: take the best k results. More is not always better.

AUGMENT

# Build a prompt with rules, evidence, and the question

| | |
|---|---|
| INSTRUCTION | Answer only from the provided context. If the context is insufficient, say so. |
| CONTEXT / 1 | [Germany Addendum §4.2] Employees may carry up to five unused days into Q1… |
| CONTEXT / 2 | [Time Off Policy §3.1] Carryover requires manager approval by December 15… |
| QUESTION | Can I carry over unused vacation days into next year? |
| OUTPUT | Give a concise answer and cite the relevant section(s). |

GENERATE + CITE

# A useful answer stays close to the evidence

HANDBOOK ASSISTANT

Yes. In Germany, you may carry over up to **five unused vacation days** into the first quarter of the next year `Germany §4.2`.

You need your manager's approval by **December 15** `Time Off §3.1`. If you have more than five days, contact People Ops for guidance.

**SPECIFIC**
Answers the employee's actual question.

**SCOPED**
Uses the Germany policy, not a global guess.

**TRACEABLE**
Every key claim points back to evidence.

WATCH THE SEAMS

# RAG can still fail

FAIL / 01

## Bad chunks

A rule is split away from its exception.

**FIX → chunk by meaning; keep overlap.**

FAIL / 02

## Wrong retrieval

The right policy never reaches the prompt.

**FIX → test queries, filters, and ranking.**

FAIL / 03

## Stale sources

An old handbook version outranks the current one.

**FIX → version, expire, and re-index.**

FAIL / 04

## Unsupported answer

The model adds a plausible detail not in context.

**FIX → require citations and abstention.**

TAKEAWAY

# RAG checklist

- [x] Prepare trusted, current sources
- [x] Chunk text without losing meaning
- [x] Embed questions and handbook chunks
- [x] Retrieve a small set of relevant evidence
- [x] Generate from context, cite, and abstain

THE MENTAL MODEL

Find the right evidence.
Give it to the model.
Check the answer.