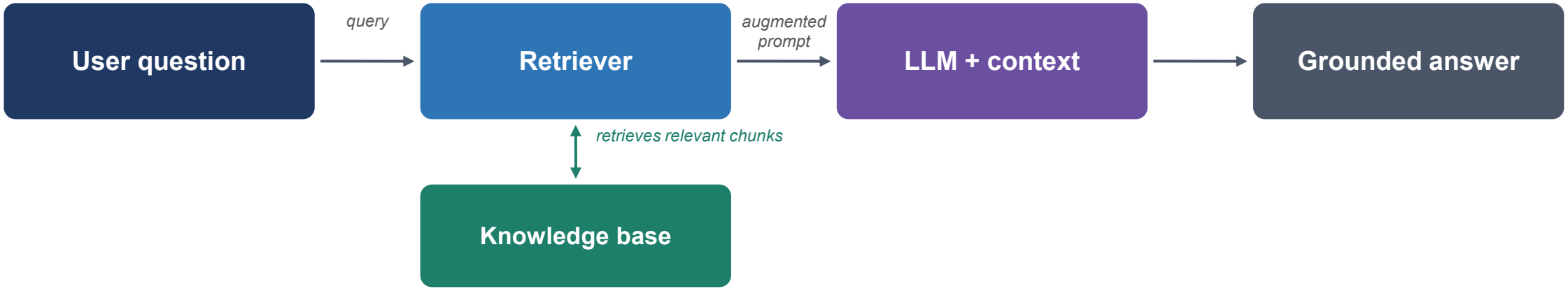
How RAG Works: Ask, Look, Answer

Before the model answers, the system retrieves relevant knowledge and gives it to the model as context.



Stage 1: Indexing

Organising the knowledge

- Documents split into chunks
- Each chunk converted to a vector
- Vectors stored in a searchable database

Stage 2: Retrieval

Finding what matters

- Question converted to a vector
- Closest-meaning chunks found
- Context assembled for the model

Stage 3: Generation

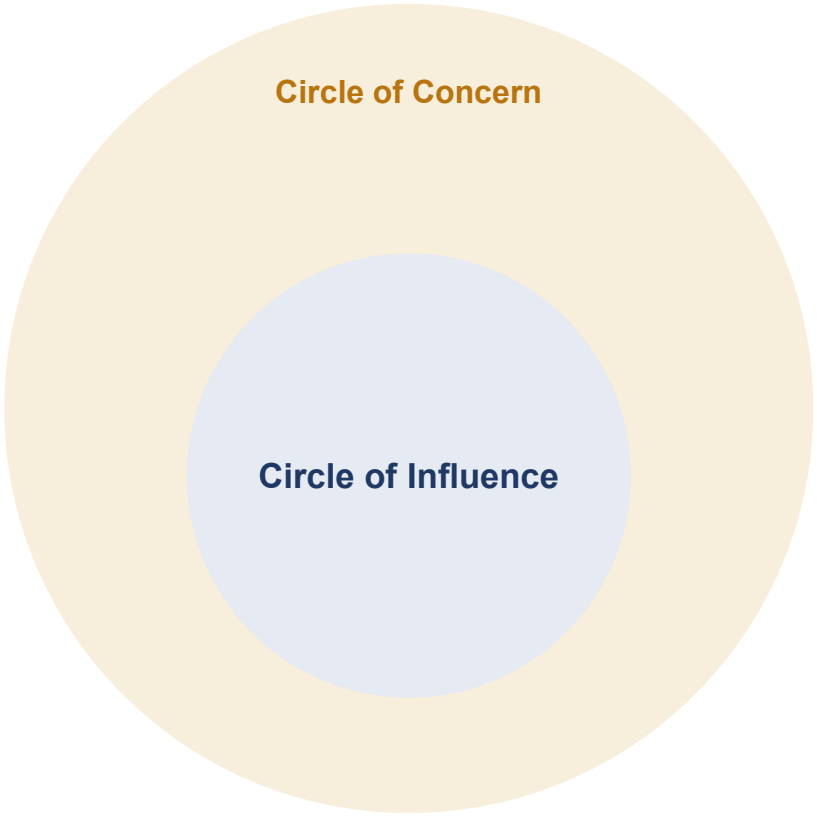
Crafting the answer

- Model receives question plus context
- Answer grounded in what was retrieved
- Retrieved first, not invented

Source: RAG: Teaching AI to Know Your Business

What You Control in a RAG System

Some things you can only worry about. Four decisions are yours, and they set the quality of the output.



You worry about it, but can't change it	You decide it
How the base model was trained	Purpose and system prompt
Its tendency to hallucinate	Knowledge base quality and scope
How the embedding model is built	Choice of LLM
The limits of the model's own knowledge	Choice of embedding model

Of the four, knowledge base quality has the highest return and is the most underestimated.

Source: [RAG: Teaching AI to Know Your Business]

Plan your RAG system

Write down the four decisions before you build. They matter more than the tool you build it in.

Purpose and system prompt

What outcome do you need, and what must the model do with what it retrieves?

Your answer

Knowledge base

What will the model read before answering? Sources, depth, and who keeps it current.

Your answer

Choice of LLM

Which model fits this task: speed for simple look-ups, reasoning for complex recommendations?

Your answer

Grounding rules

How will you keep answers inside the retrieved context, and what happens when it is not enough?

Your answer

Source: [RAG: Teaching AI to Know Your Business]

About this diagram

RAG Architecture

When this fits

- Anyone planning an AI assistant that must answer from their own knowledge: policies, products, documentation or expertise.
- Explaining to stakeholders why the knowledge base, not the model, sets the quality ceiling.
- Diagnosing a RAG system that gives generic or inconsistent answers.

When it doesn't fit

- Tasks that need the AI to take actions in other systems. That needs a connection layer such as MCP.
- Open-ended creative work where there is no body of knowledge to retrieve from.

How to use this file

- Every element is a native PowerPoint shape: edit the text, recolour it, or delete what you don't need.
- Use the blank template slide (the one headed Map your or Plan your) to fill in your own version.
- Copy the diagram straight into your own deck. It is sized for a standard widescreen slide.

Source: RAG: Teaching AI to Know Your Business