Microsoft

Generative AI with JavaScript

Improve AI accuracy and reliability with RAG

Known AI challenges

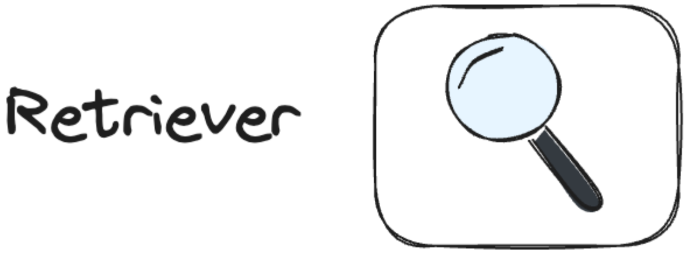
- Wrong or inaccurate information
- Out of date data
- Non-authoritative sources

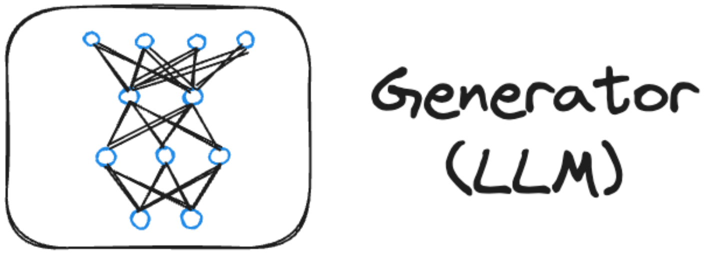
Retrieval-Augmented Generation (RAG)

- Combination of a retriever and a generator

Retrieval-Augmented Generation (RAG)

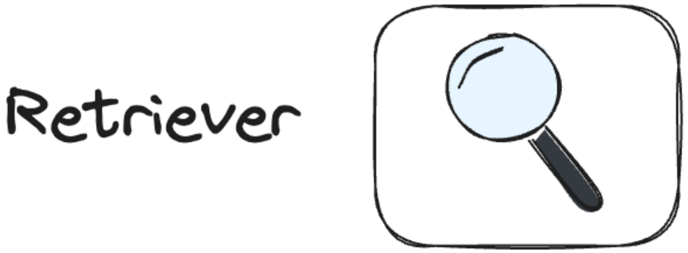
- Combination of a retriever and a generator

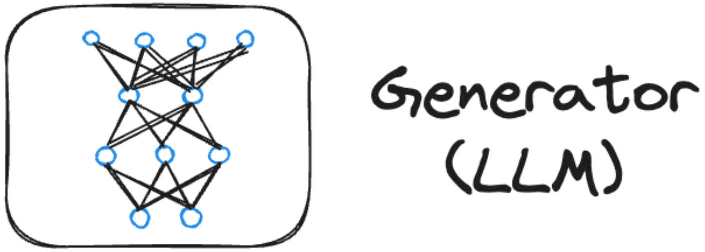




Retrieval-Augmented Generation (RAG)

- Combination of a retriever and a generator
- Allows for more precise and relevant answers

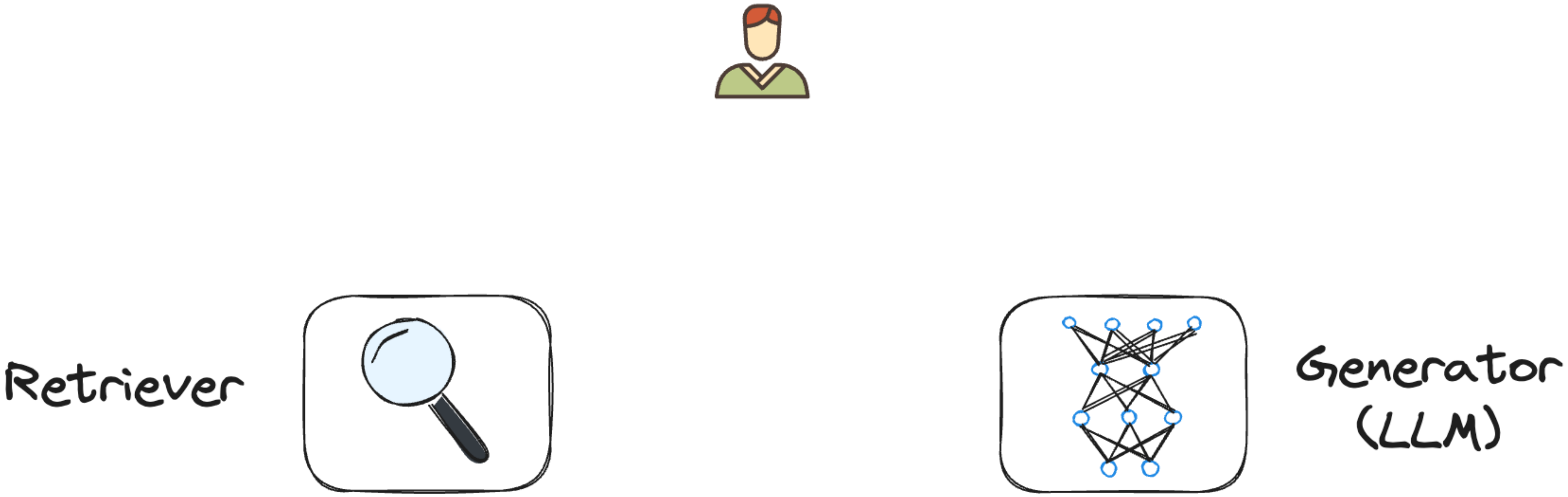




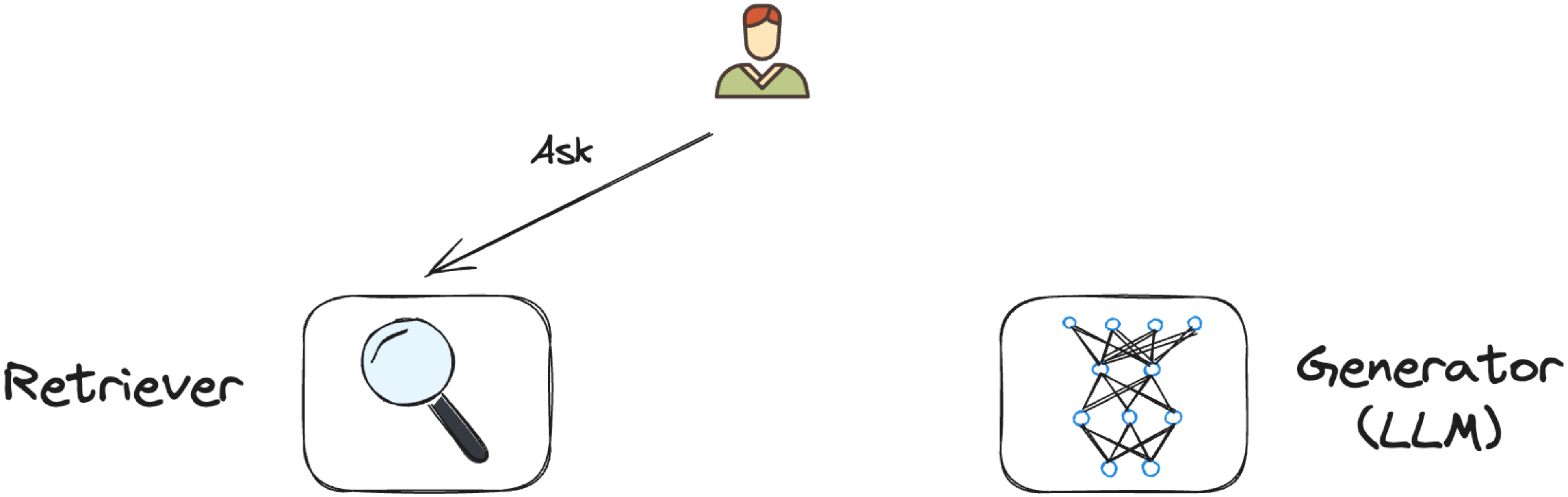
Why RAG?

- No training needed
- Up-to-date data
- Groundedness
- User trust

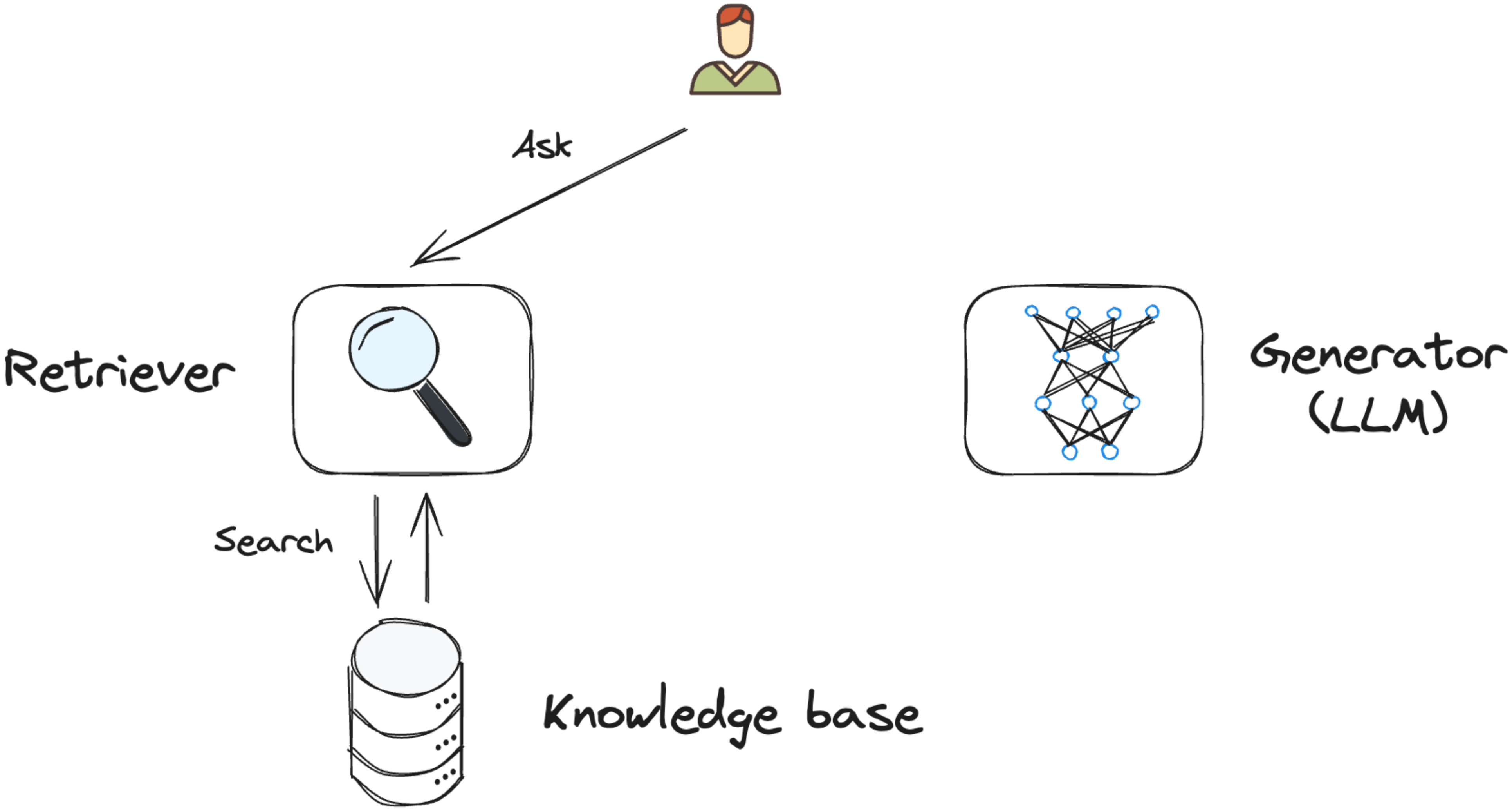
Retrieval-Augmented Generation (RAG)



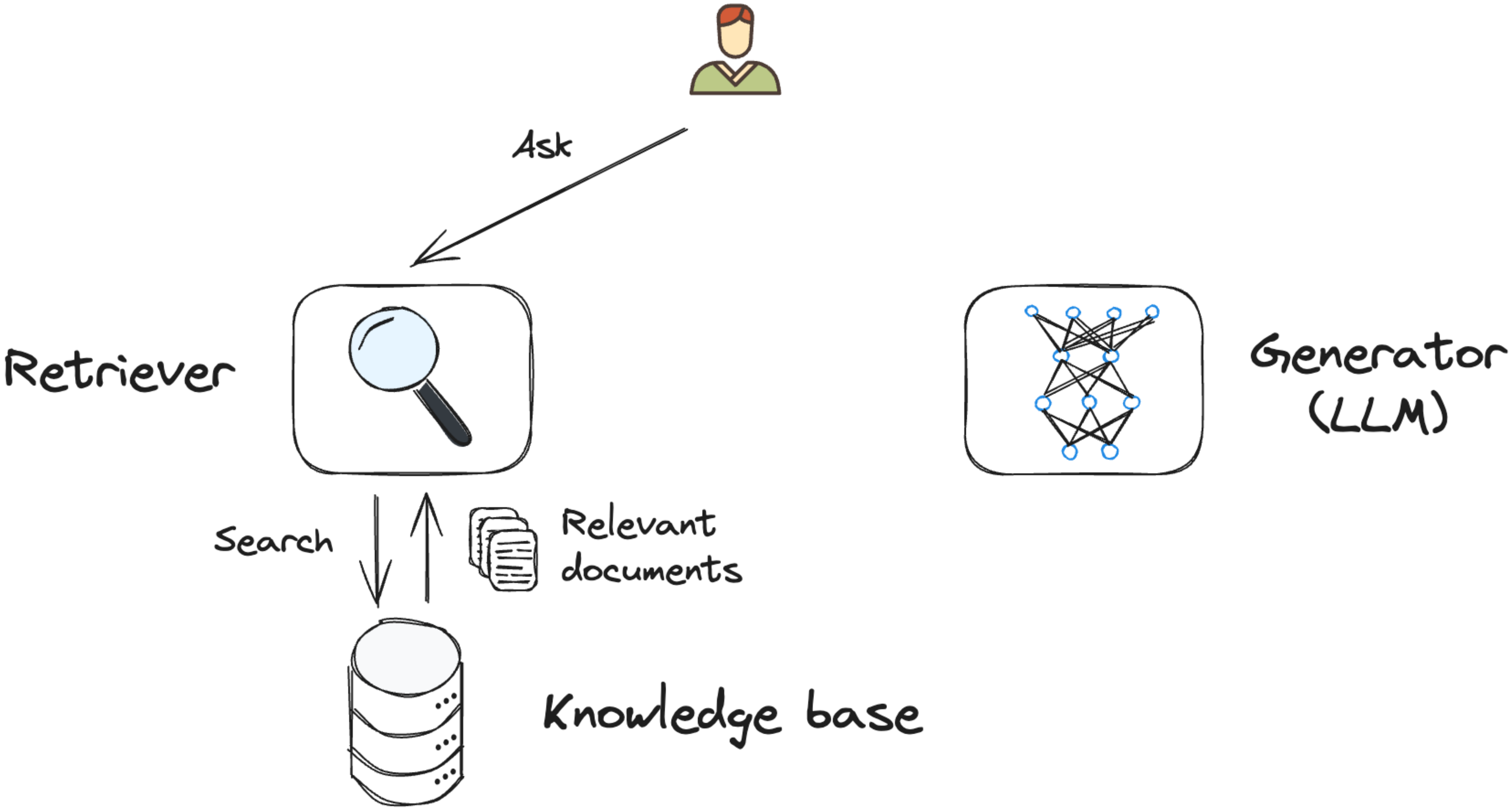
Retrieval-Augmented Generation (RAG)



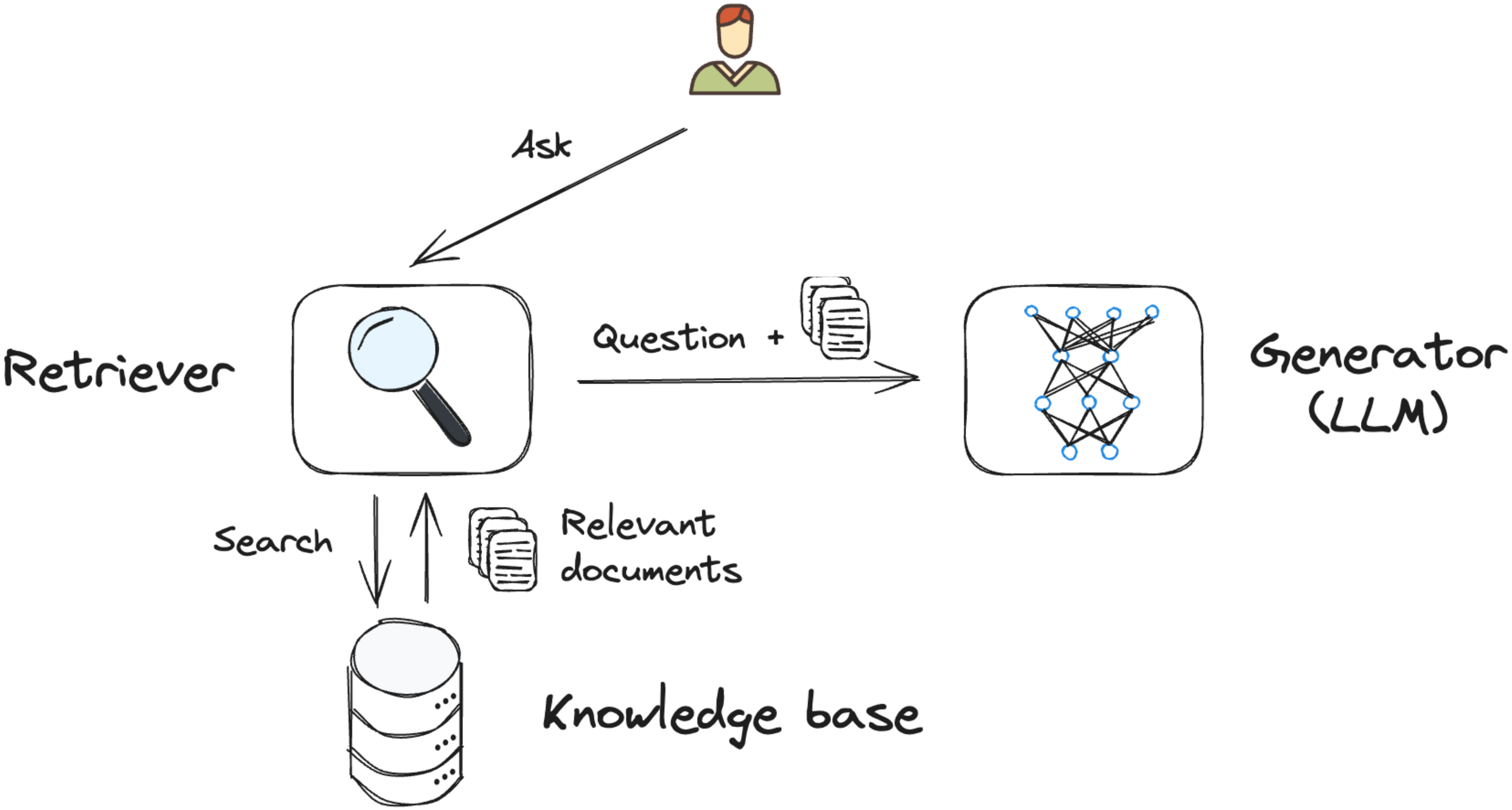
Retrieval-Augmented Generation (RAG)



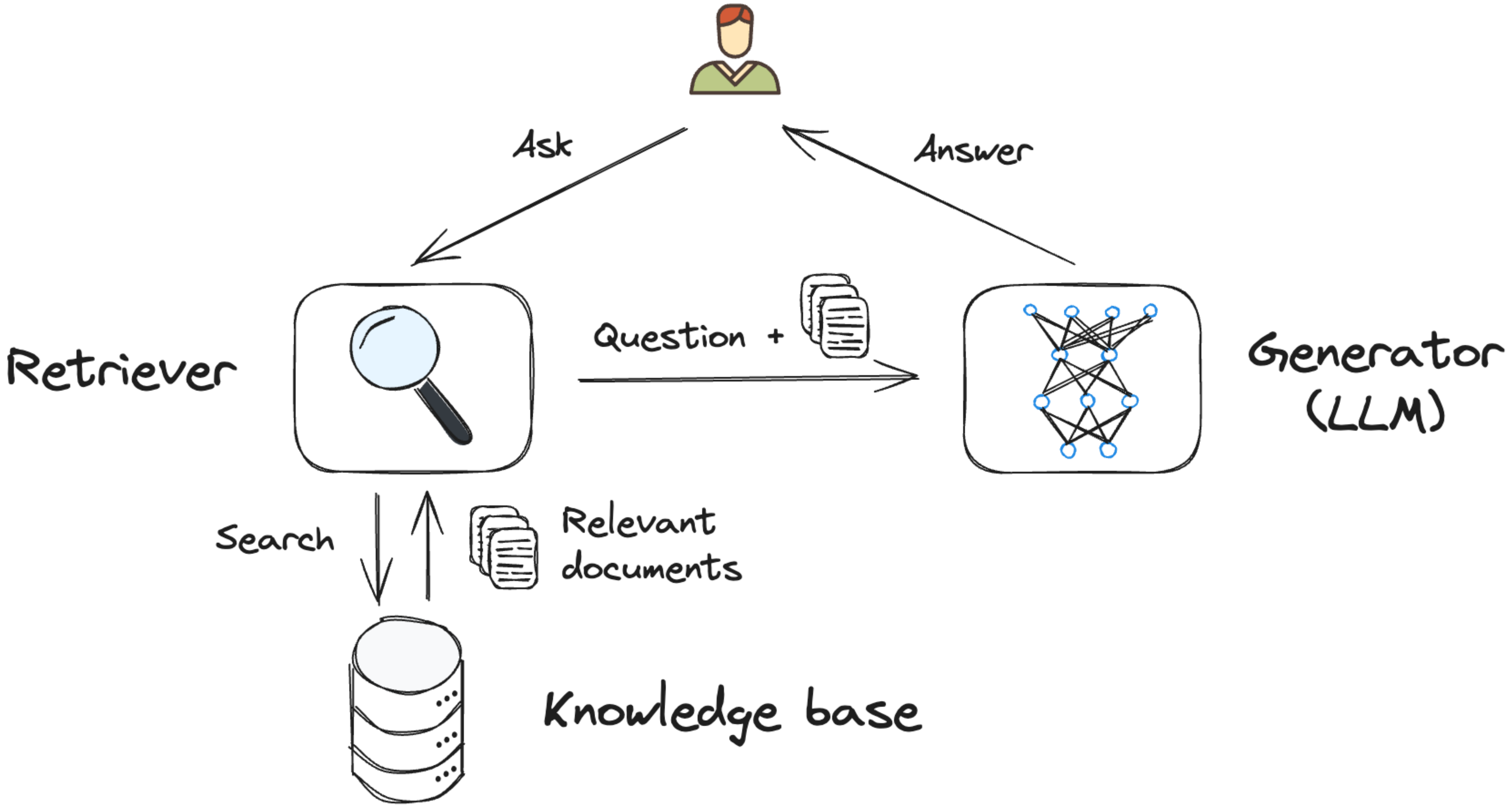
Retrieval-Augmented Generation (RAG)



Retrieval-Augmented Generation (RAG)



Retrieval-Augmented Generation (RAG)

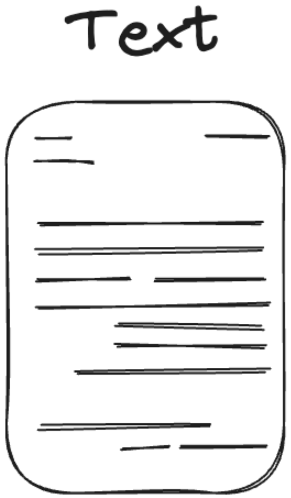


Building the knowledgebase

Create document embeddings from text data to use in the retriever

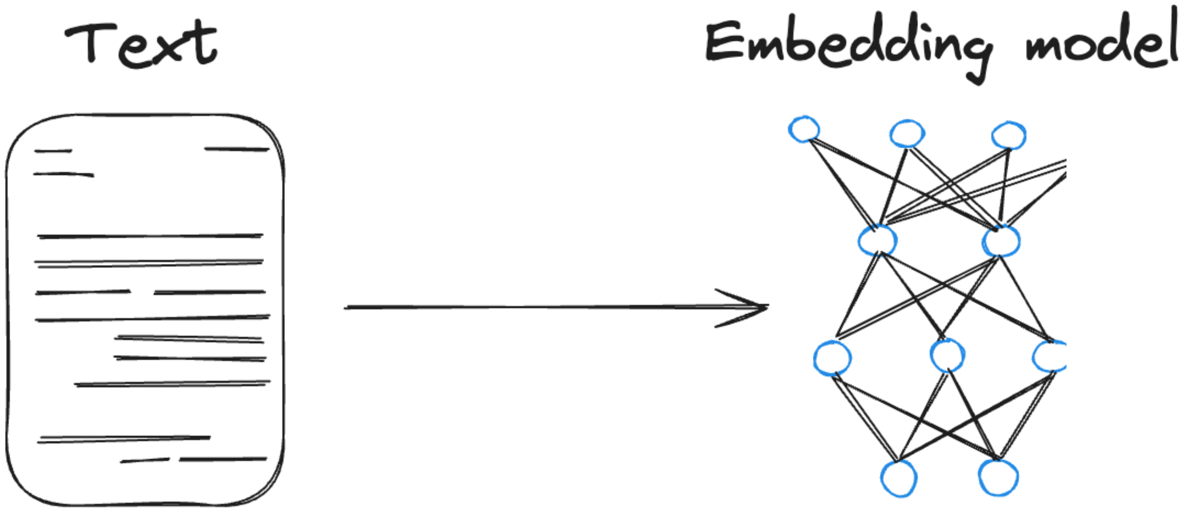
Building the knowledgebase

Create document embeddings from text data to use in the retriever



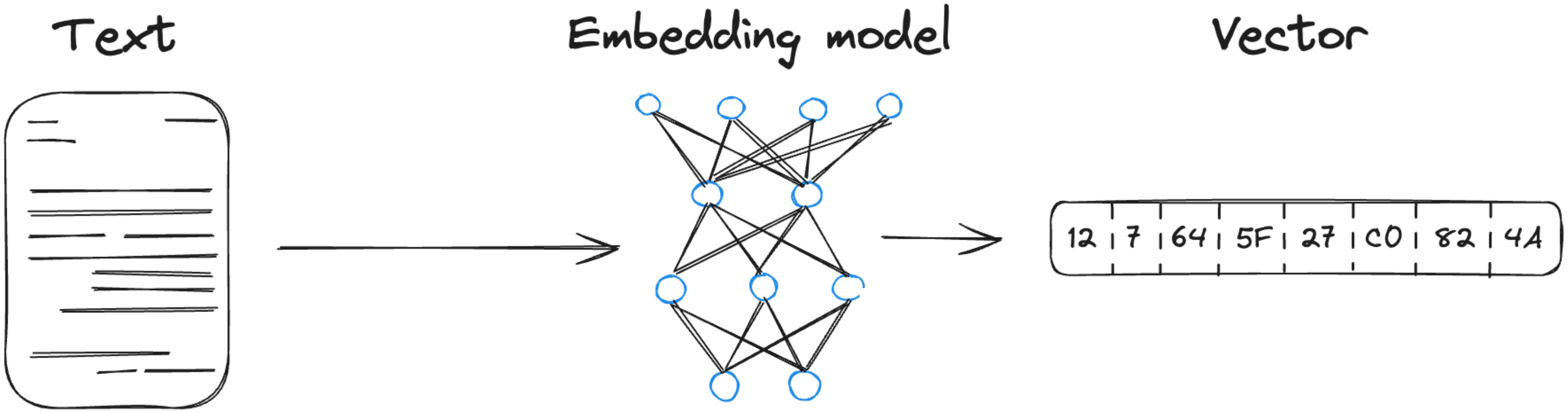
Building the knowledgebase

Create document embeddings from text data to use in the retriever

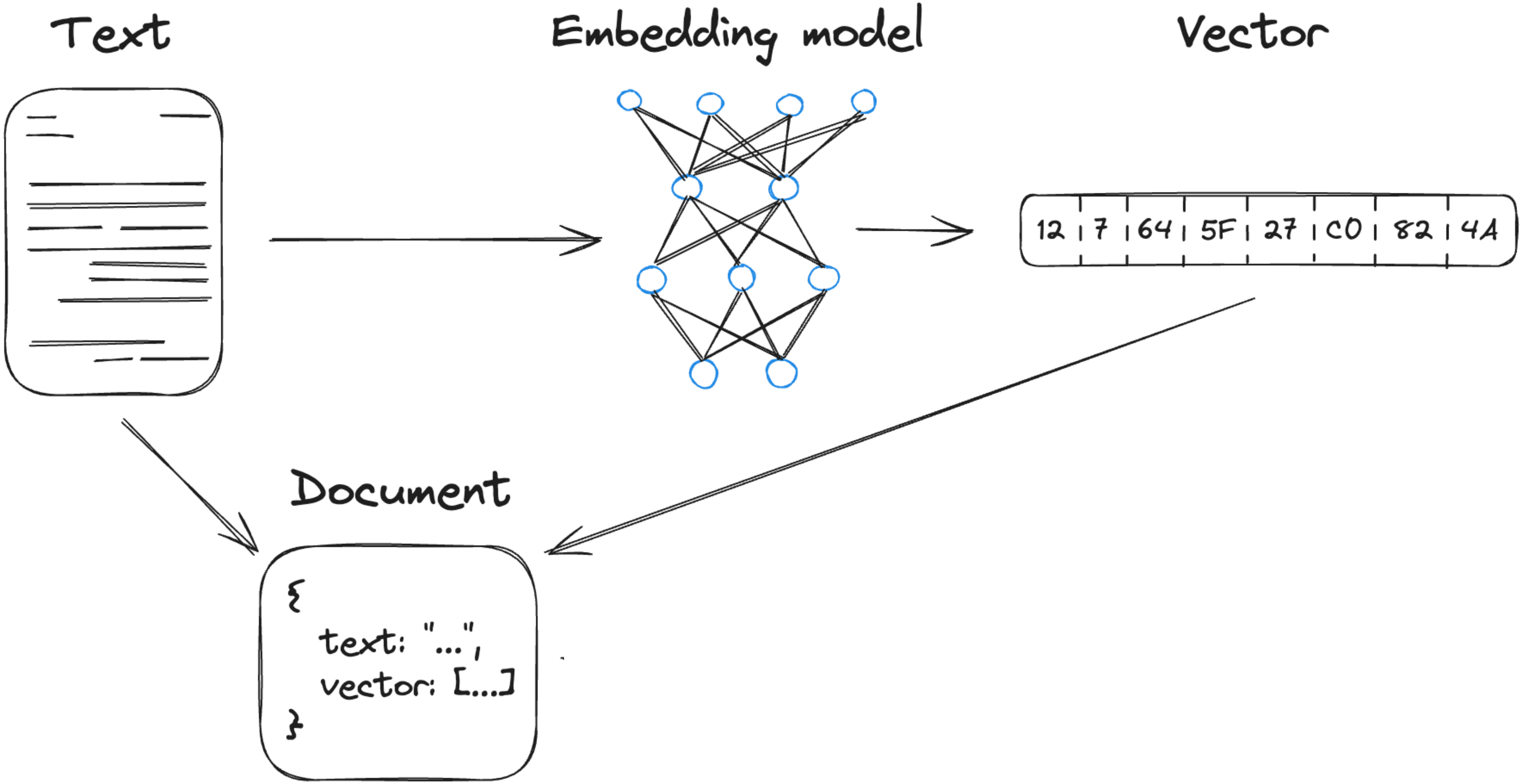


Building the knowledgebase

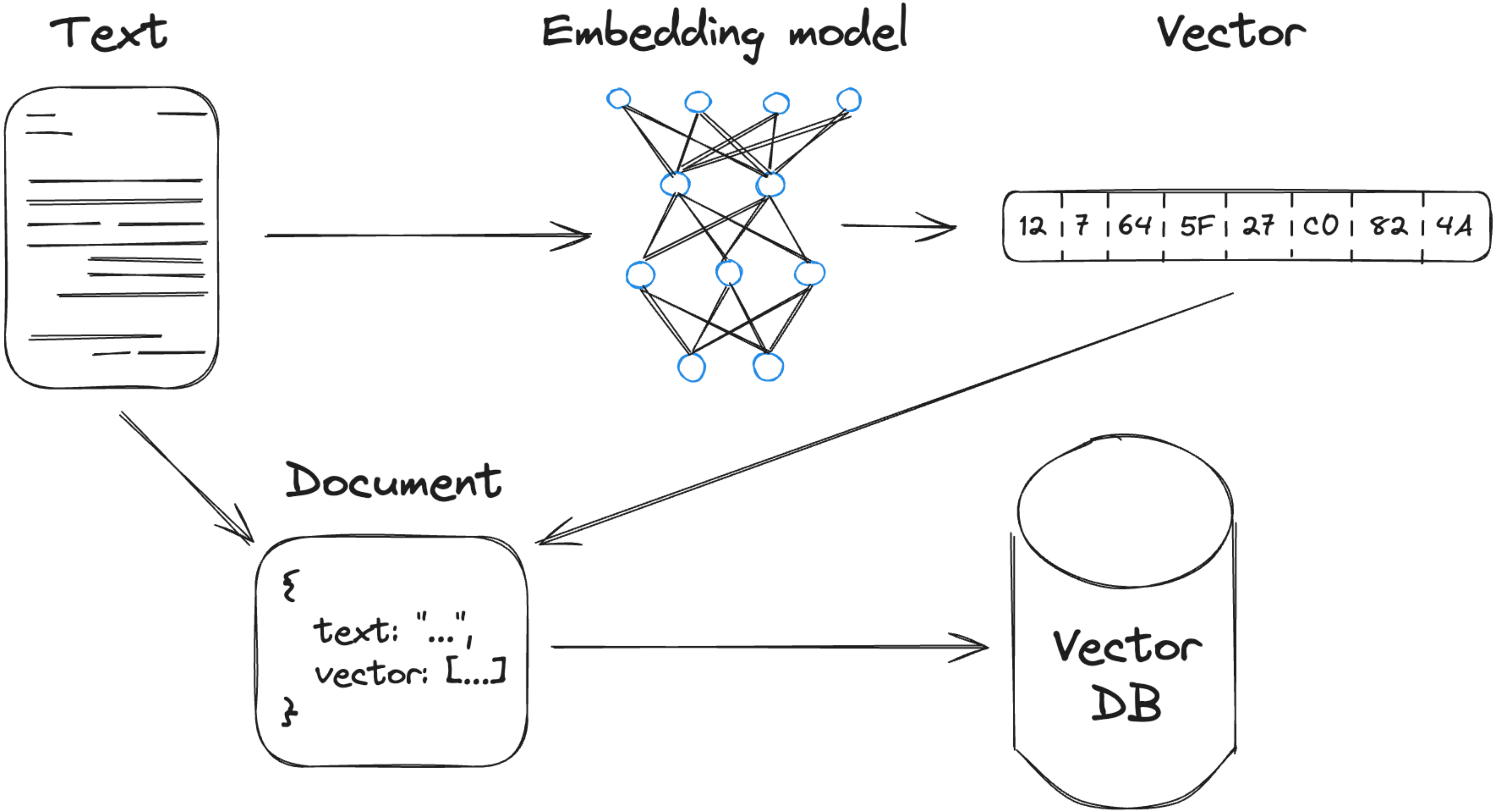
Create document embeddings from text data to use in the retriever



Building the knowledgebase



Building the knowledgebase



Retrieval and context augmentation

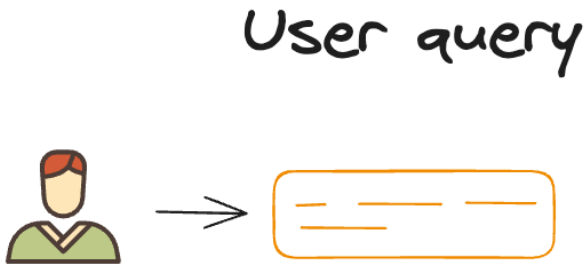
1. Transform the user query into a vector
2. Search in vector DB for relevant documents
3. Add found documents to the context

Retrieval and context augmentation

1. Transform the user query into a vector

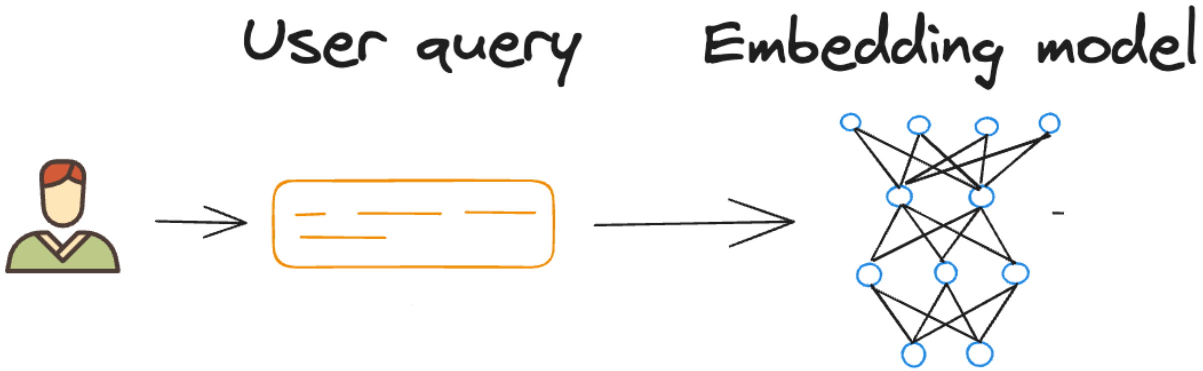
Retrieval and context augmentation

1. Transform the user query into a vector



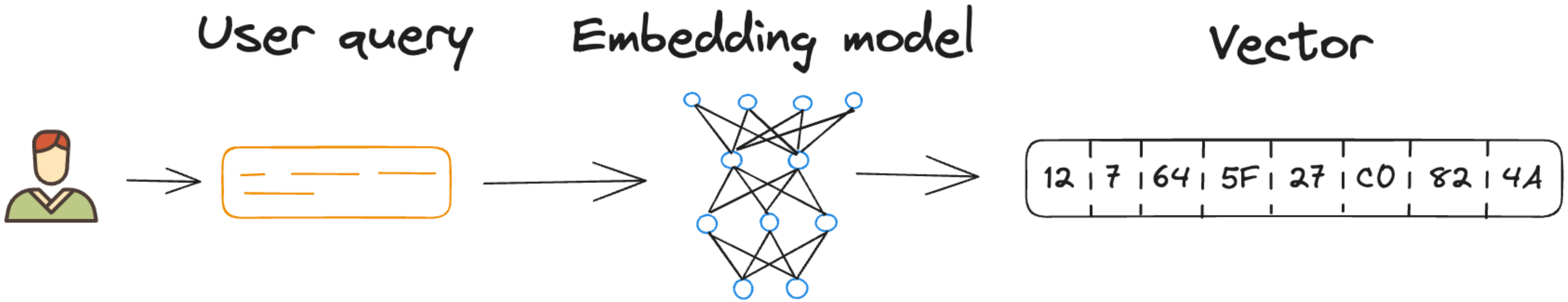
Retrieval and context augmentation

1. Transform the user query into a vector



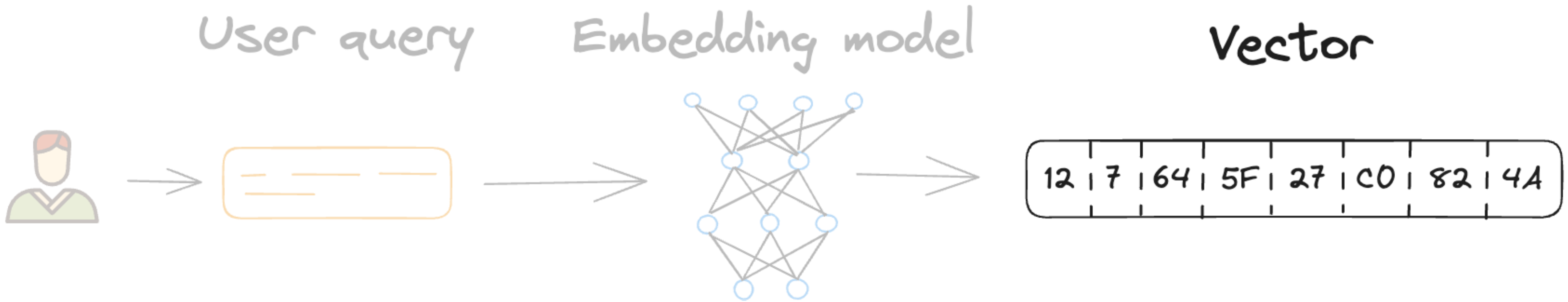
Retrieval and context augmentation

1. Transform the user query into a vector



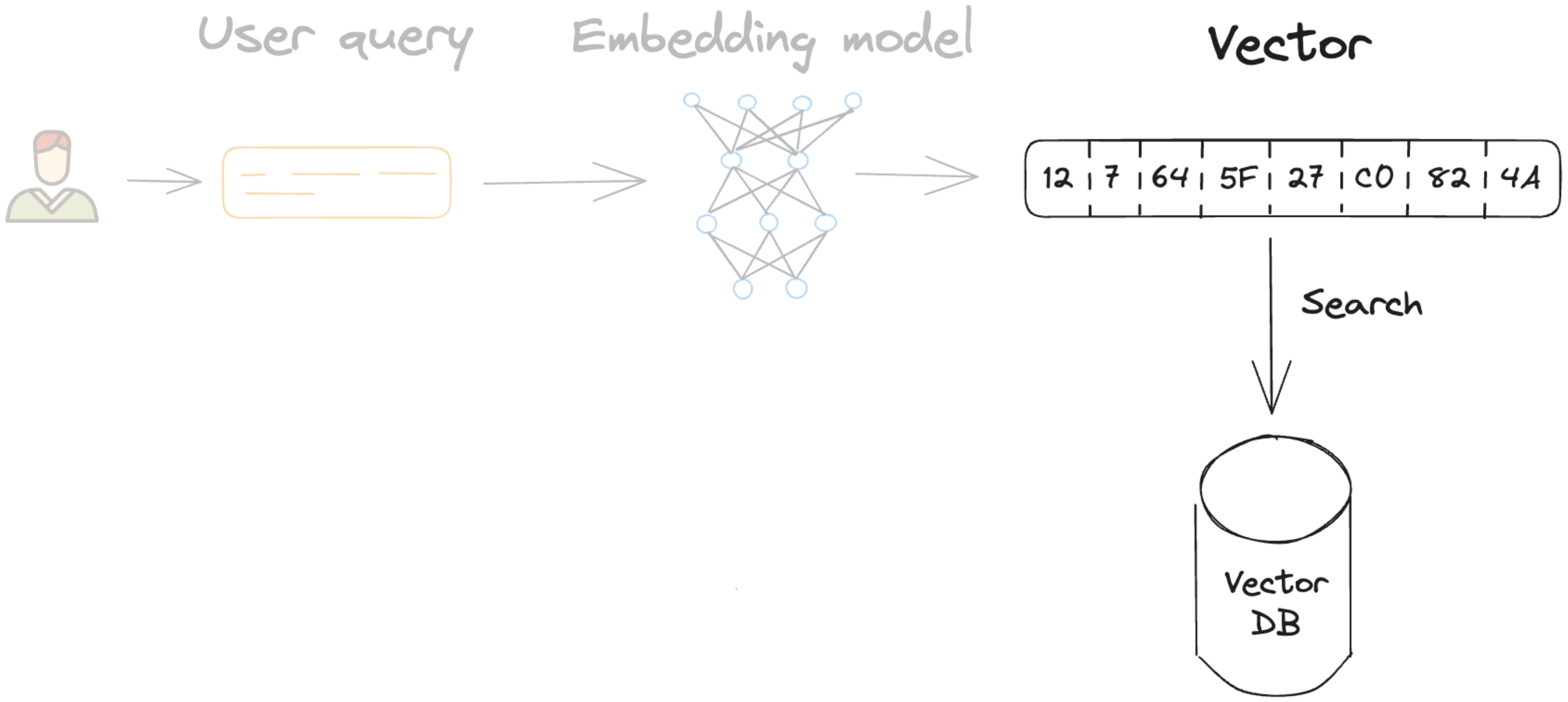
Retrieval and context augmentation

2. Search in vector DB for relevant documents



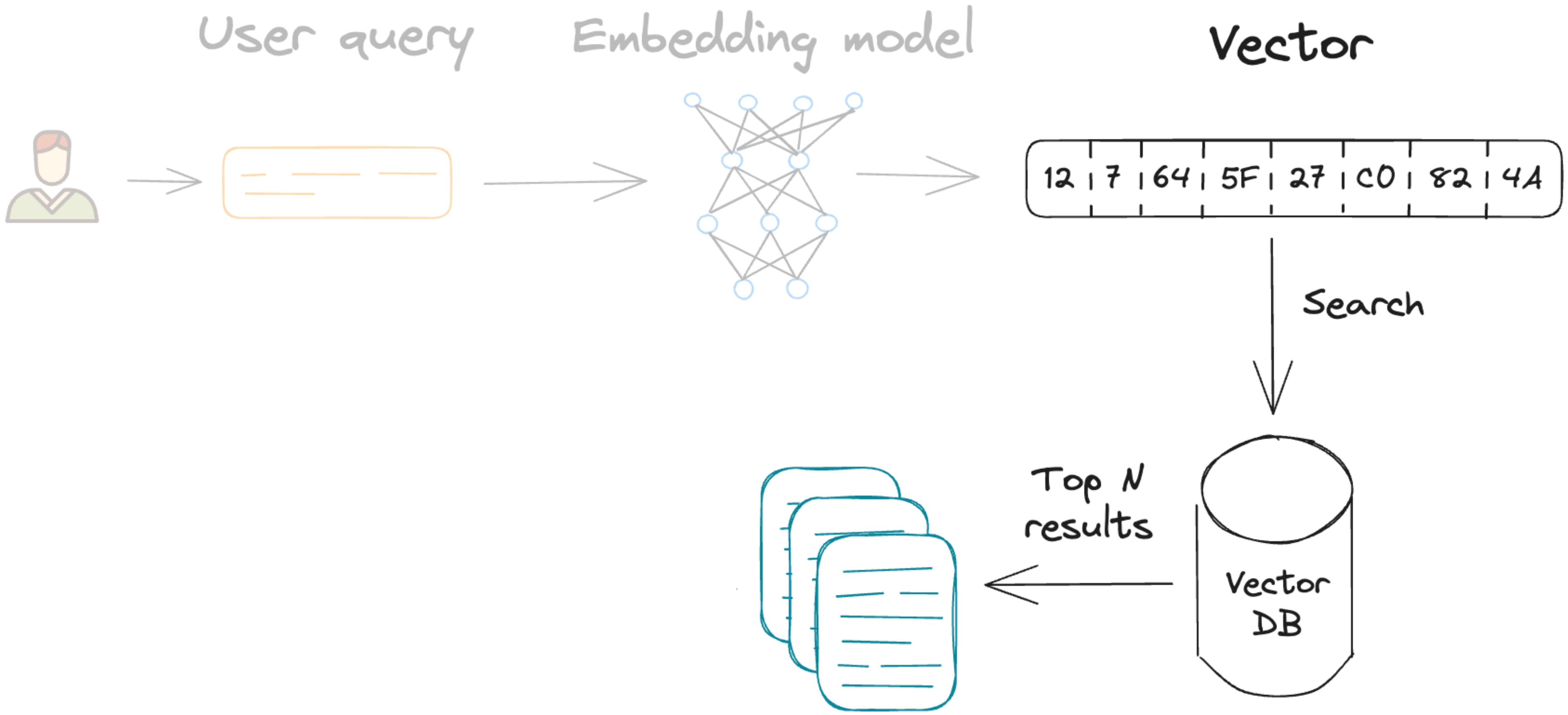
Retrieval and context augmentation

2. Search in vector DB for relevant documents



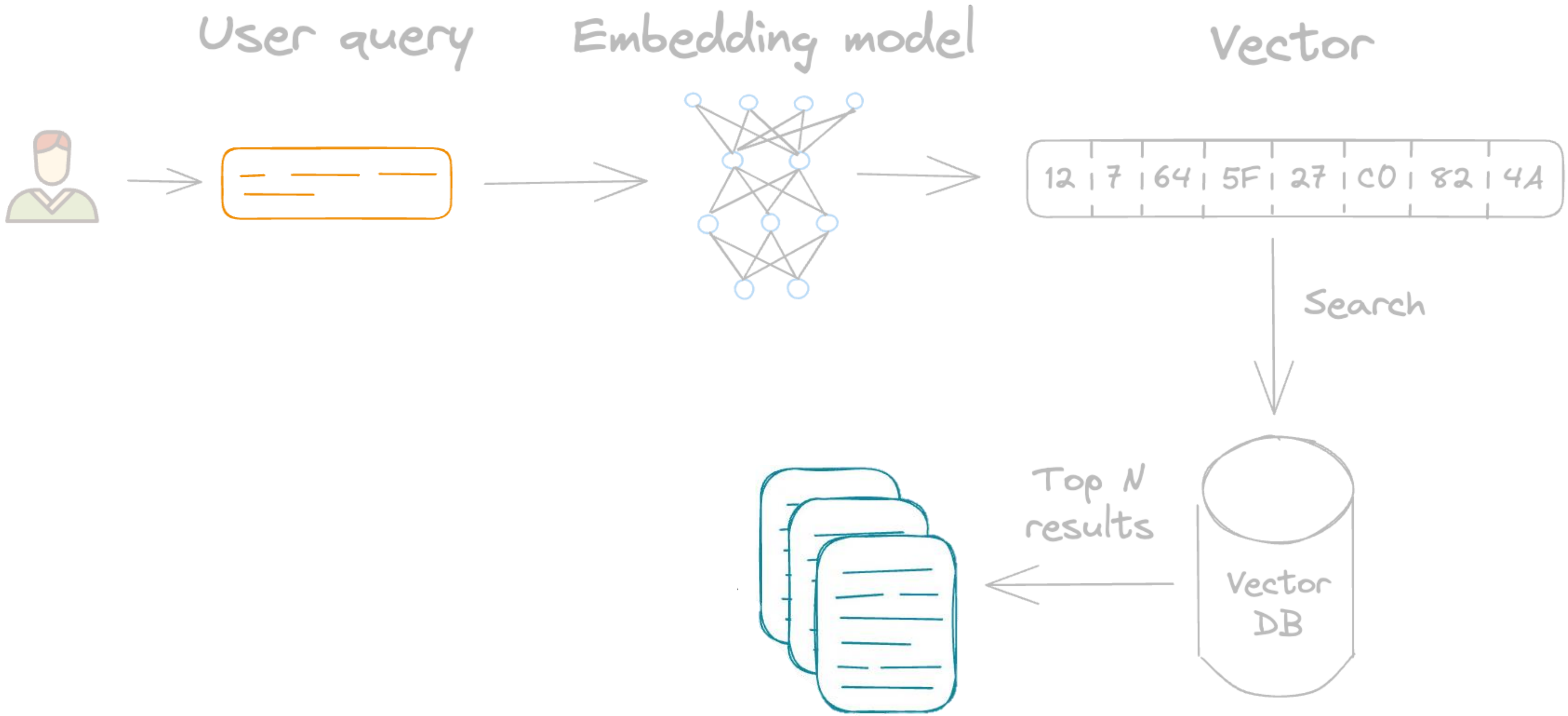
Retrieval and context augmentation

2. Search in vector DB for relevant documents



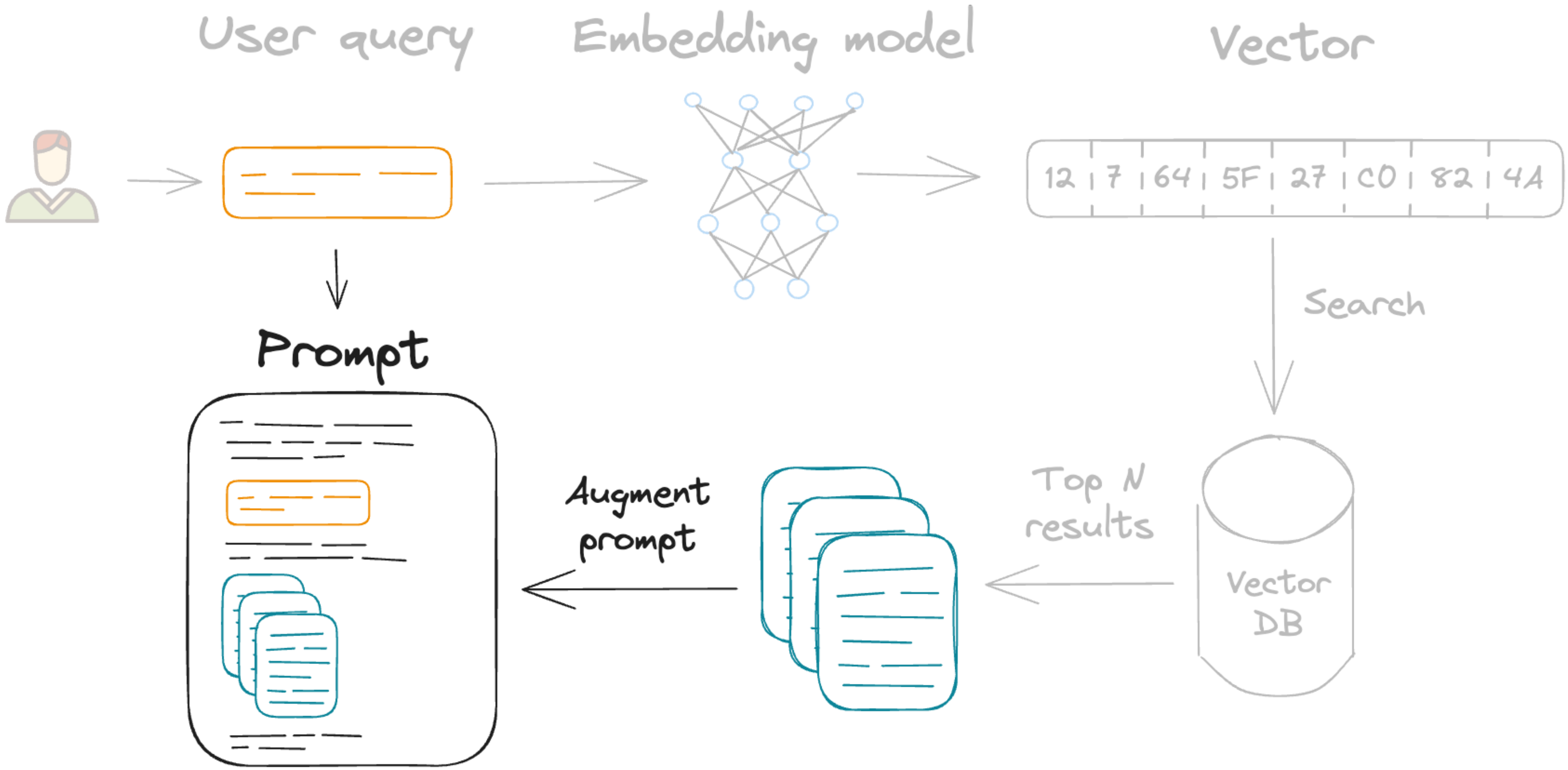
Retrieval and context augmentation

3. Add found documents to the context



Retrieval and context augmentation

3. Add found documents to the context



Company support chat with RAG

Demo – aka.ms/ai/js/chat

Microsoft

Resources

• Serverless AI chat demo	aka.ms/ai/js/chat
• Implement RAG training	aka.ms/genai/rag
• AI chat with RAG workshop	aka.ms/ws/openai-rag-qdrant

Microsoft