Skills Network

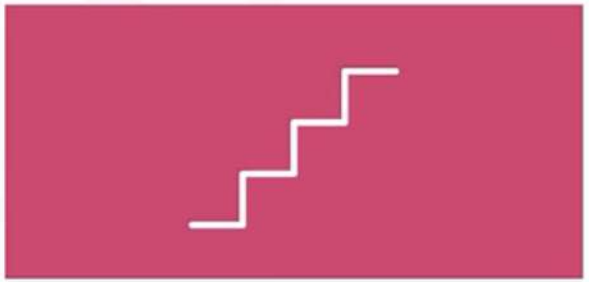
Retrieval-Augmented Generation (RAG)

IBM

Retrieval Augmented Generation (RAG)

What you will learn

Explain the RAG process

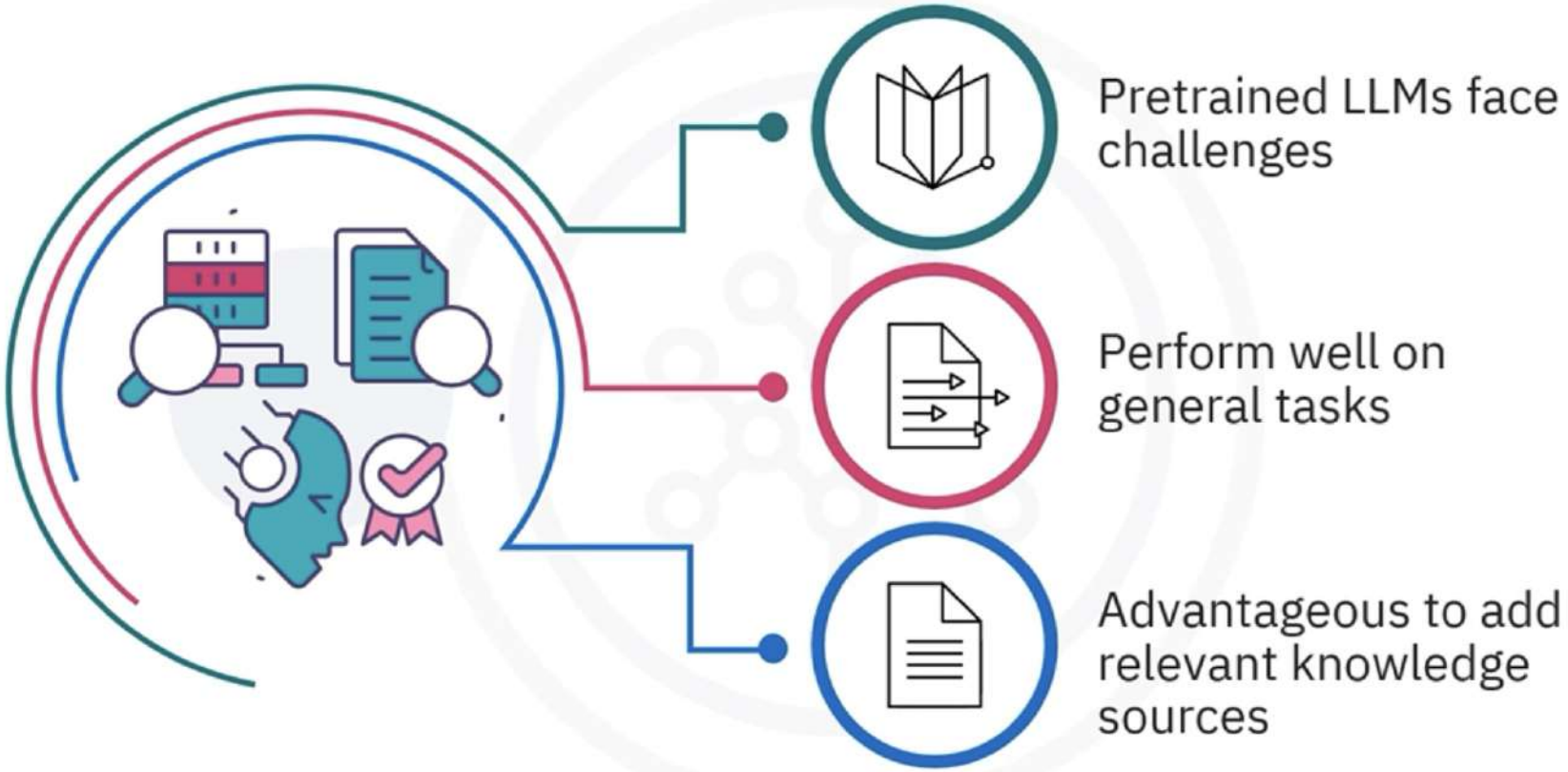


Describe various steps in the RAG process

What is RAG?


An AI framework
Optimizes the output of LLM
Uses LLMs' capabilities

Importance of RAG in training LLMs



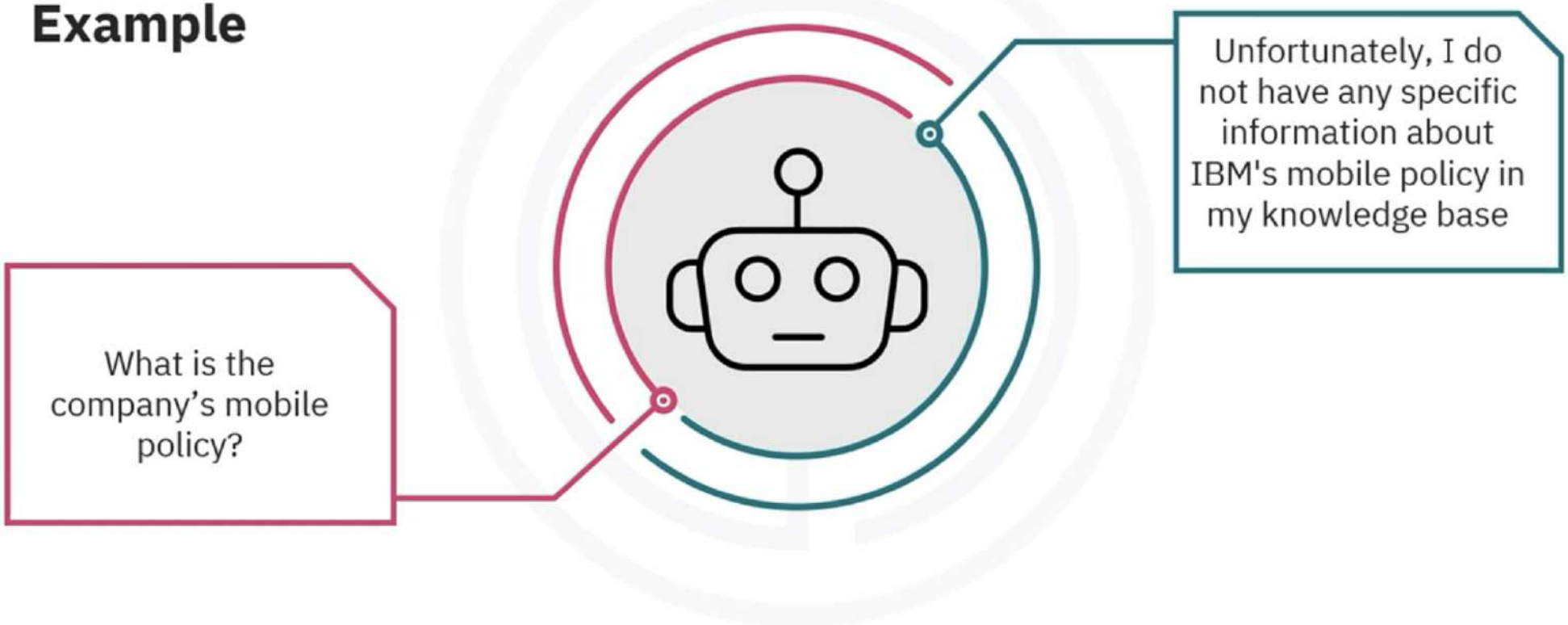
Pretrained LLMs face challenges

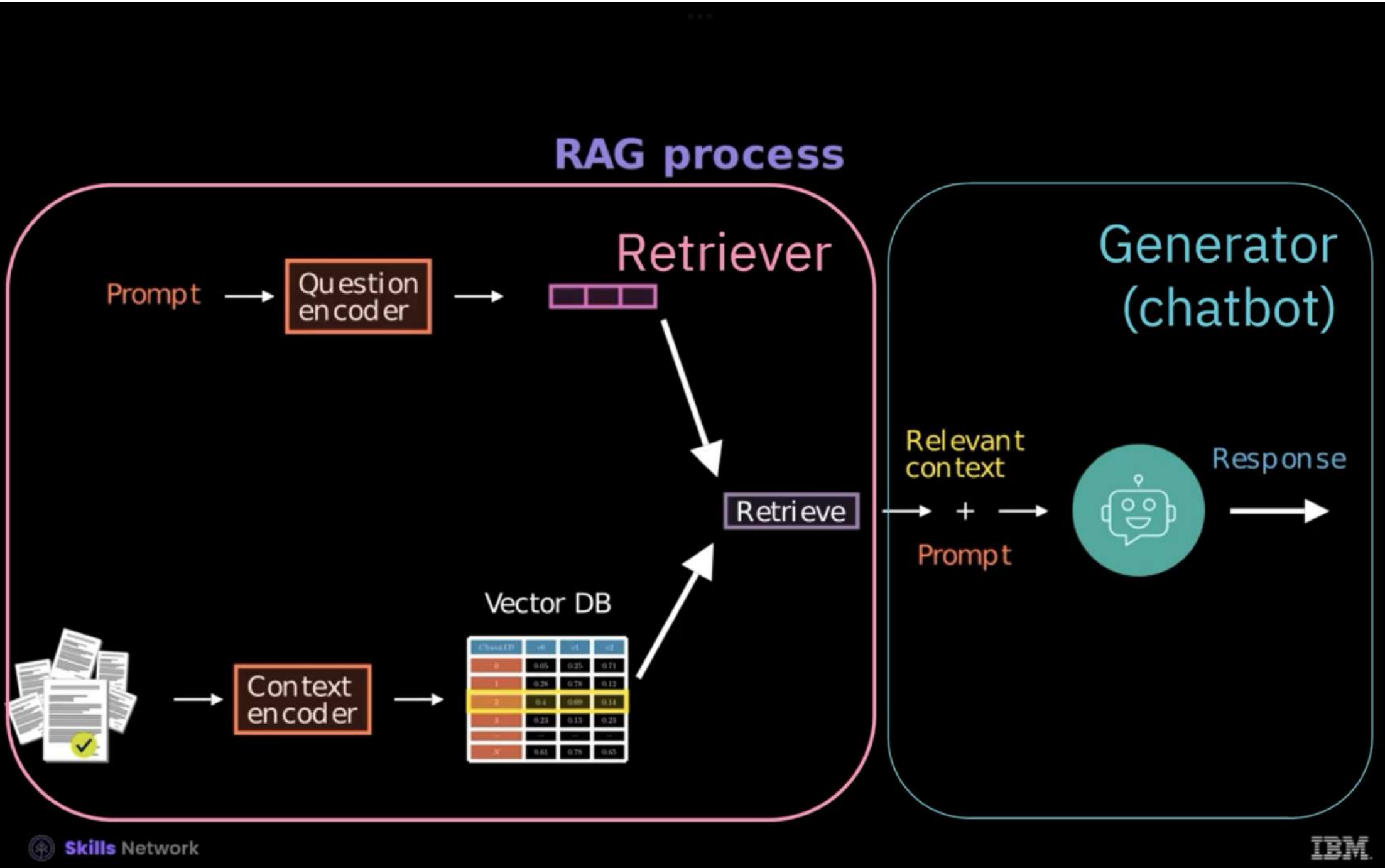
Perform well on general tasks

Advantageous to add relevant knowledge sources

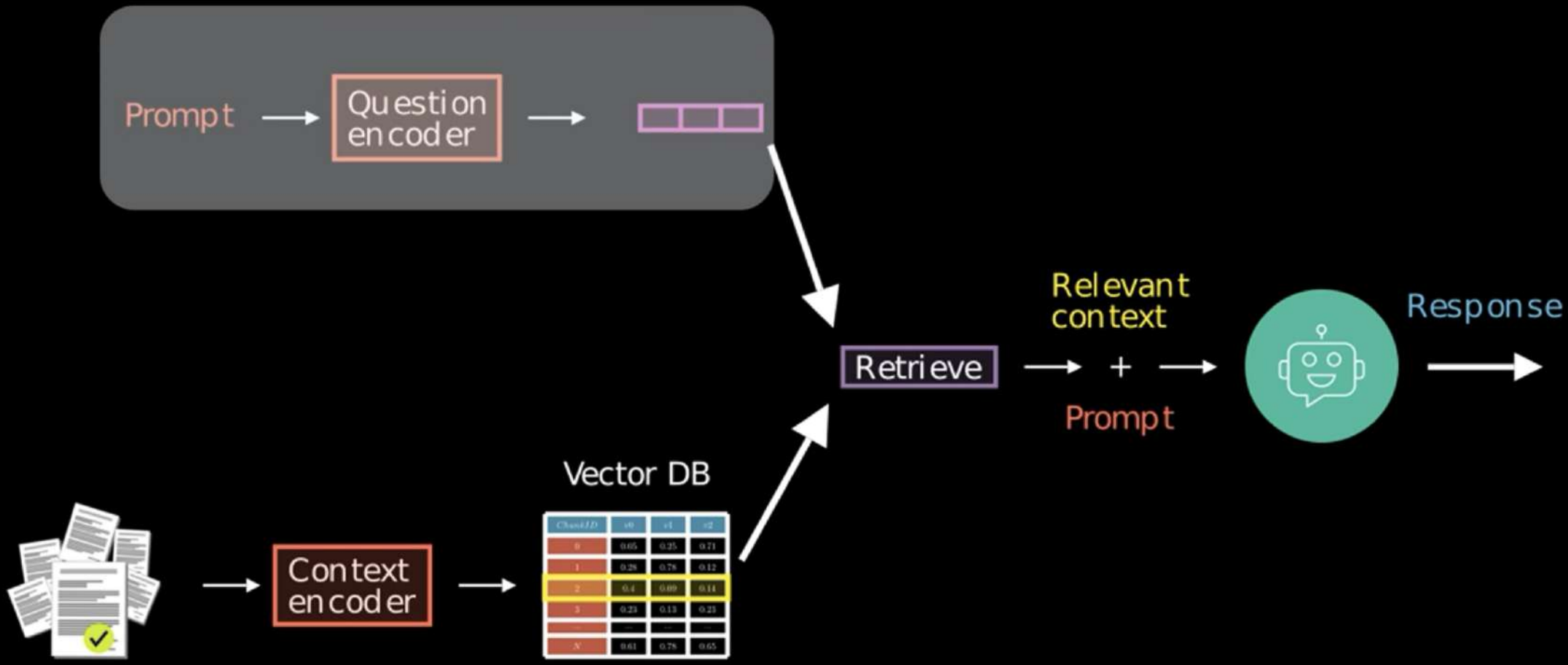
Importance of RAG in training LLMs

Example



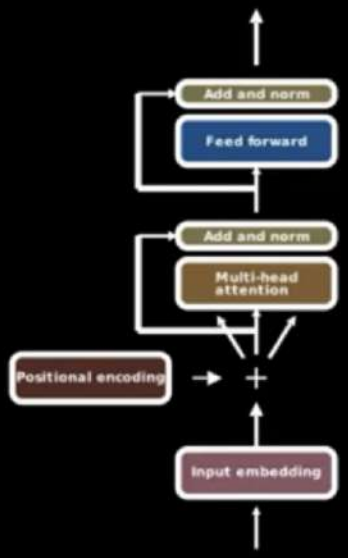

RAG process
Retriever
Prompt
Question encoder
Retrieve
Vector DB
Context encoder
Generator (chatbot)
Relevant context
+
Prompt
Response
Skills Network
IBM

RAG process


Prompt
Question encoder
Retrieve
Relevant context
+
Prompt
Response
Vector DB
Context encoder

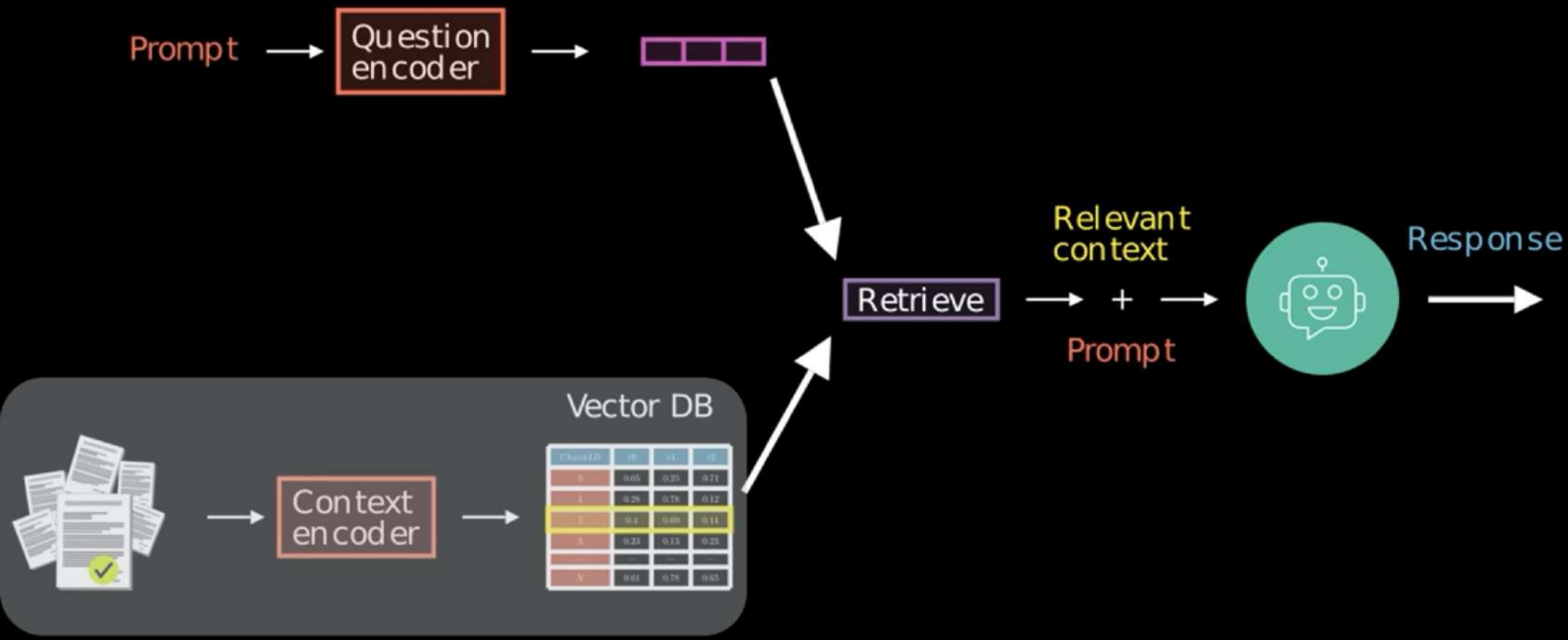
Questions to vectors

$$\frac{1}{N}\sum\Big(\quad \cdots \quad \Big) =$$



question:
What is your mobile policy?

RAG process


Prompt
Question encoder
Relevant context
Retrieve
+
Prompt
Response
Vector DB
Context encoder

Context encoding

Company policy on mobile security...

Policy on employee conduct is the...

Our mobile policy allows employees...

Health and safety are impact ...

Equal opportunities are key in...

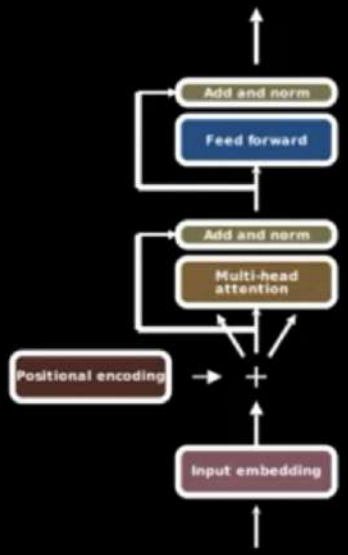
Company environmental impact in ...

Company policy on mobile devices...

Context encoding

Chunks to vectors

$$\frac{1}{N}\sum\left(\quad \cdots \quad \right) =$$



Chunk 2:
Our mobile policy allows employees to use personal devices for work.

Context encoding

Documents

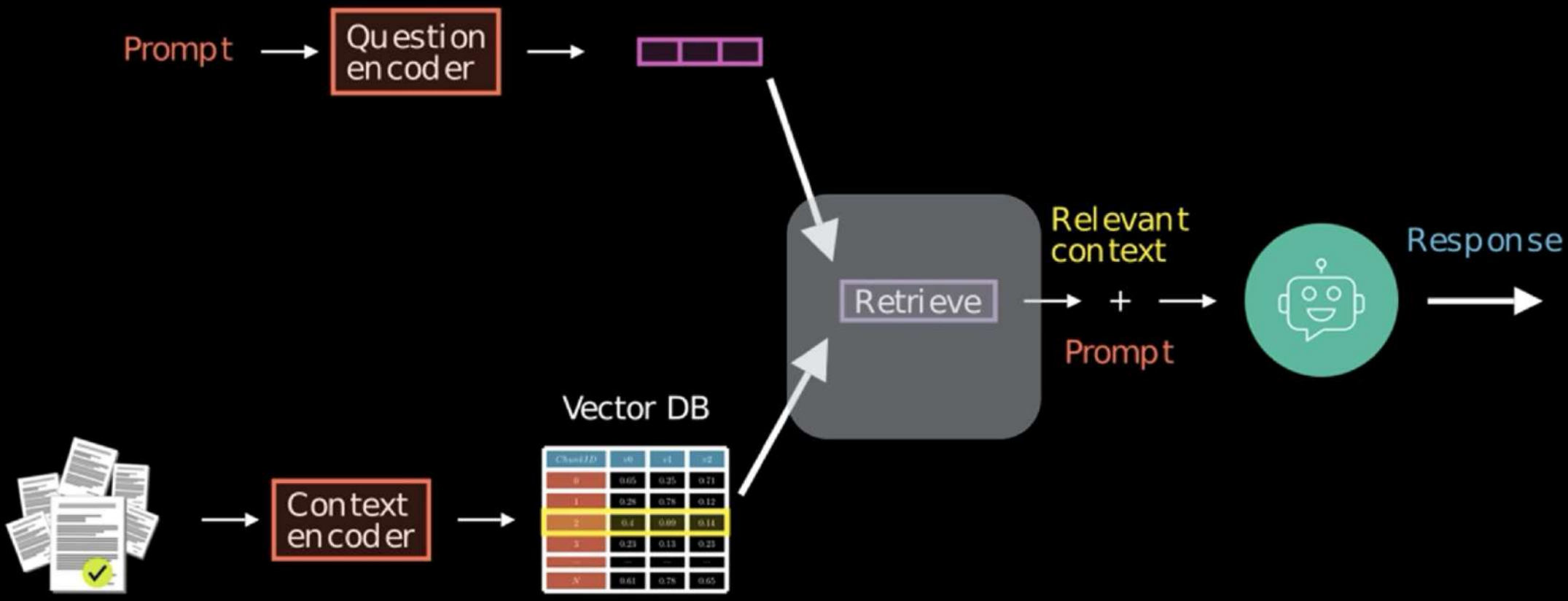
Chunk 0
Chunk 1
Chunk 2
Chunk 3
Chunk 4
Chunk 5
Chunk 6

Embedding vectors

Knowledge base

ChunkID	h_1	h_2	h_3
0	0.05	0.25	0.71
1	0.28	0.78	0.12
2	0.4	0.09	0.14
3	0.23	0.13	0.23
...	...	...	...
N	0.61	0.78	0.65

RAG process


Prompt
Question encoder
Retrieve
Relevant context
+
Prompt
Response
Vector DB
Context encoder

Search relevant context

Question: What is your mobile policy?

Question vector

0.35	0.08	0.16

Knowledge base

ChunkID	h_1	h_2	h_3	*Distance*
0	0.05	0.25	0.71	0.65
1	0.28	0.78	0.12	0.71
2	0.4	0.09	0.14	0.05
3	0.23	0.13	0.23	0.89
...	...	...	...	...
N	0.61	0.78	0.65	0.15

→

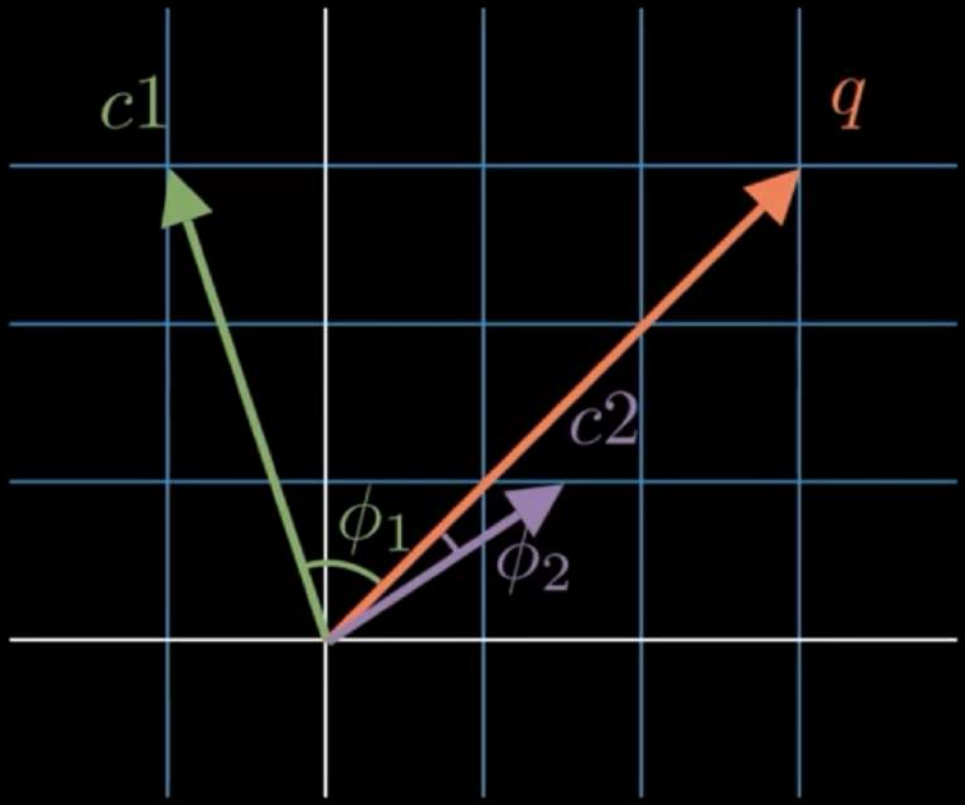
Relevant context:
Our mobile policy allows employees to use personal devices for work.

Vector similarity

Dot product

$$q \cdot c = \sum (q_i \times c_i)$$

$$\Rightarrow q \cdot c1 > q \cdot c2$$



Cosine similarity

$$\cos(\phi) = \frac{q \cdot c}{\|q\| \times \|c\|}$$

$$\Rightarrow \cos(\phi_1) < \cos(\phi_2)$$

Which of the two context vectors c1 and c2 is more similar to question vector q?

Select top K relevant context

Question: What is your mobile policy?

Question vector

0.35	0.08	0.16

Knowledge base

ChunkID	h_1	h_2	h_3	*Distance*
0	0.05	0.25	0.71	0.65
1	0.28	0.78	0.12	0.71
2	0.4	0.09	0.14	0.05
3	0.23	0.13	0.23	0.89
...	...	...	...	...
N	0.61	0.78	0.65	0.15

Select top K relevant context

ChunkID	*Distance*
0	0.65
1	0.71
2	0.05
3	0.89
...	...
$N-1$	0.15

Select top K relevant context

$$K = 3$$

$$\delta = \text{argsort}_{i<K} \begin{pmatrix} 0.65 & 0.71 & 0.05 & 0.89 & \dots & 0.15 \end{pmatrix}$$

0.65	0.71	0.05	0.89	···	0.15
0	1	2	3	···	$N-1$

Select top K relevant context

$$K = 3$$

$$\delta = [2, 6, 0]$$

Relevant context

Chunk 0

Chunk 1

Chunk 2

Chunk 3

Chunk 4

Chunk 5

Chunk 6

Relevant context

Chunk 2

Chunk 6

Chunk 0

Relevant context

Our mobile policy allows employees...

Company policy on mobile devices...

Company policy on mobile security...

Response generation

Question: What is your mobile policy?

Our mobile policy allows employees...

Company policy on mobile devices...

Company policy on mobile security...

Recap

- RAG helps generate responses
- Challenging for chatbot to generate responses for specific domains such as the company's mobile policy
- To generate responses, a chatbot:
  - Encodes inserted question or prompt
  - Breaks down into smaller chunks of text
  - Converts text chunks into high-dimensional vectors using distance metrics
  - Selects a vector closer to the text chunks from the knowledge base to generate a relevant response