



Vector Embeddings and RAG Demystified

Veliswa Boya

Senior Developer Advocate, AWS

A brief conversation ~~with~~ about generative AI

The capital of France is ...

Generative AI – Base models

The capital of France is Paris.

User

AI

Generative AI – Base models

The capital of France is Paris.

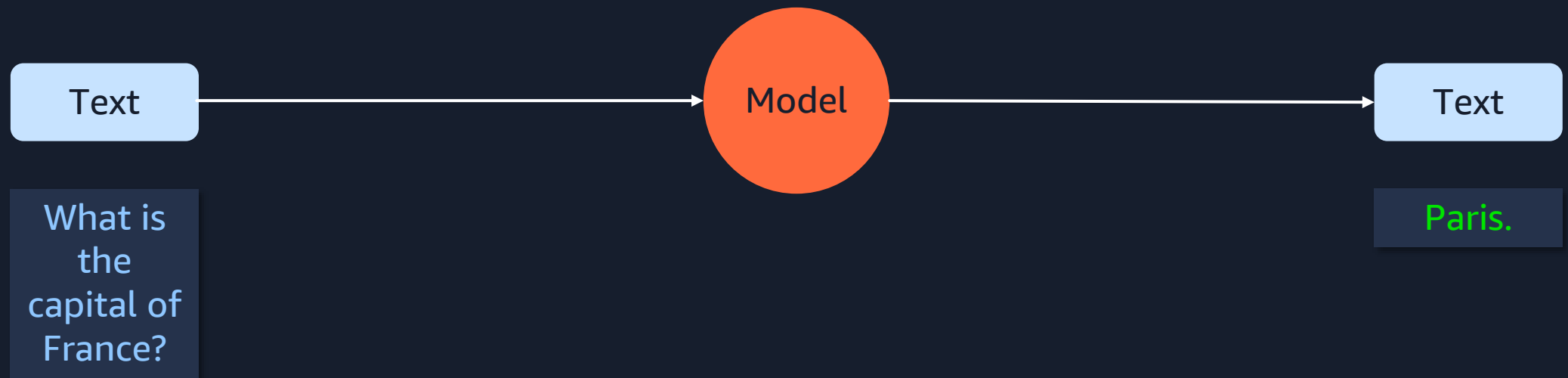
User

AI

At the end of the day, LLM simply generates one token at a time

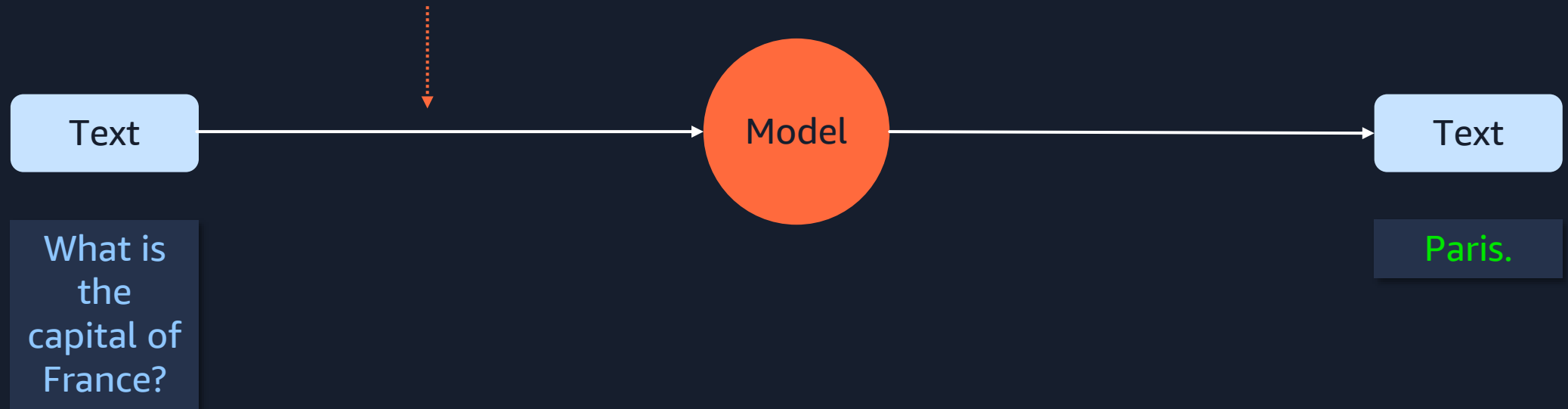
But how it does that ?

How do **large language models (LLMs)** work?



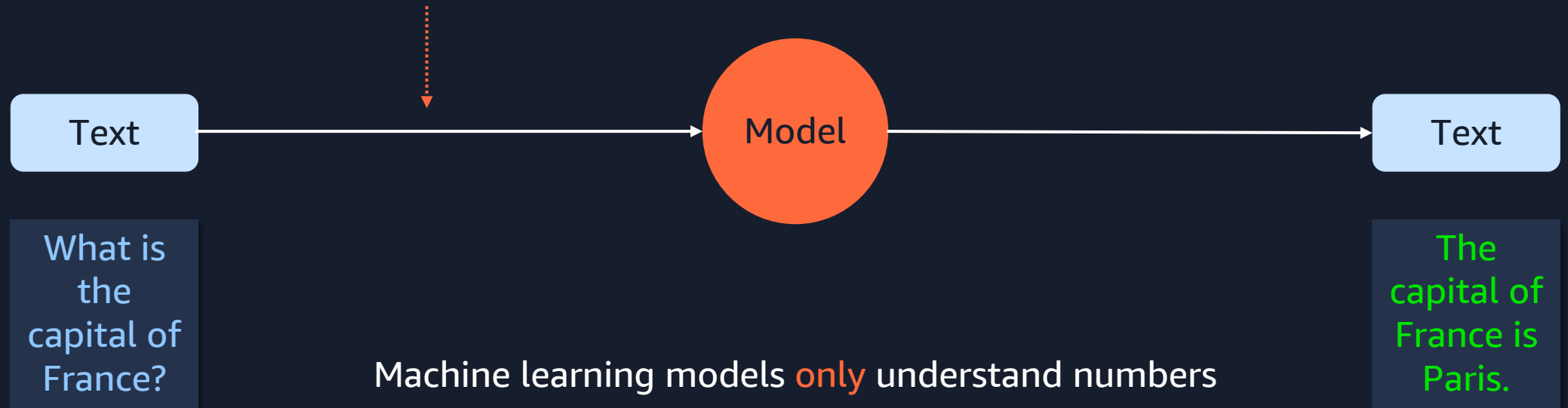
How do **large language models (LLMs)** work?

Therefore we **can not** send raw text directly



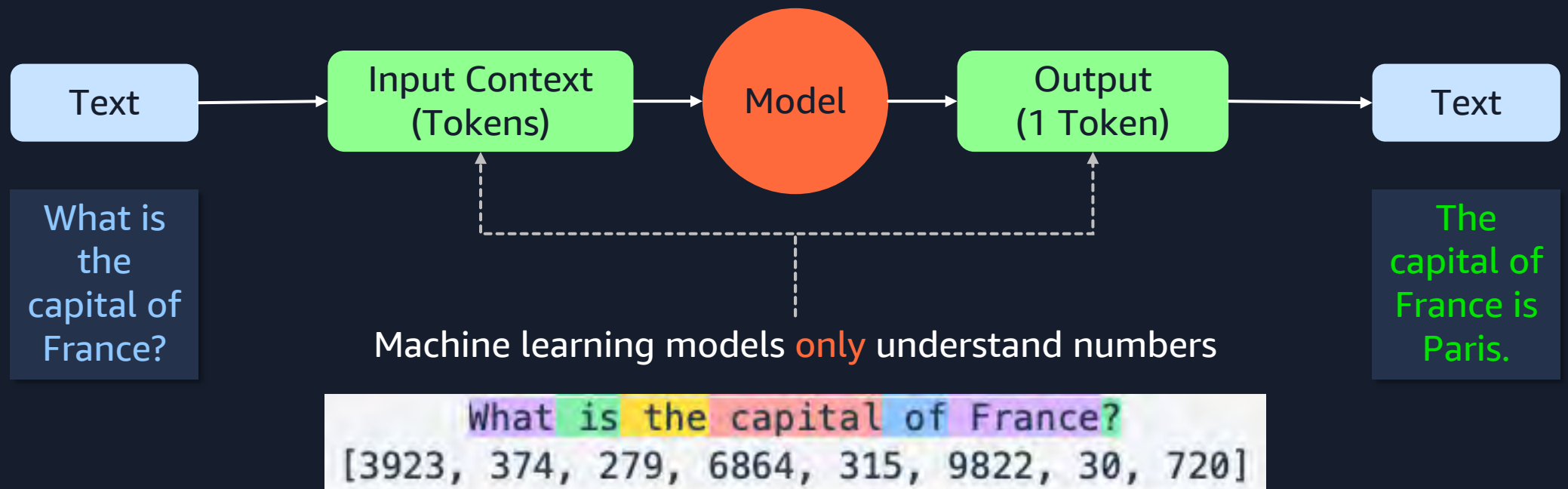
How do **large language models (LLMs)** work?

Therefore we **can not** send raw text directly

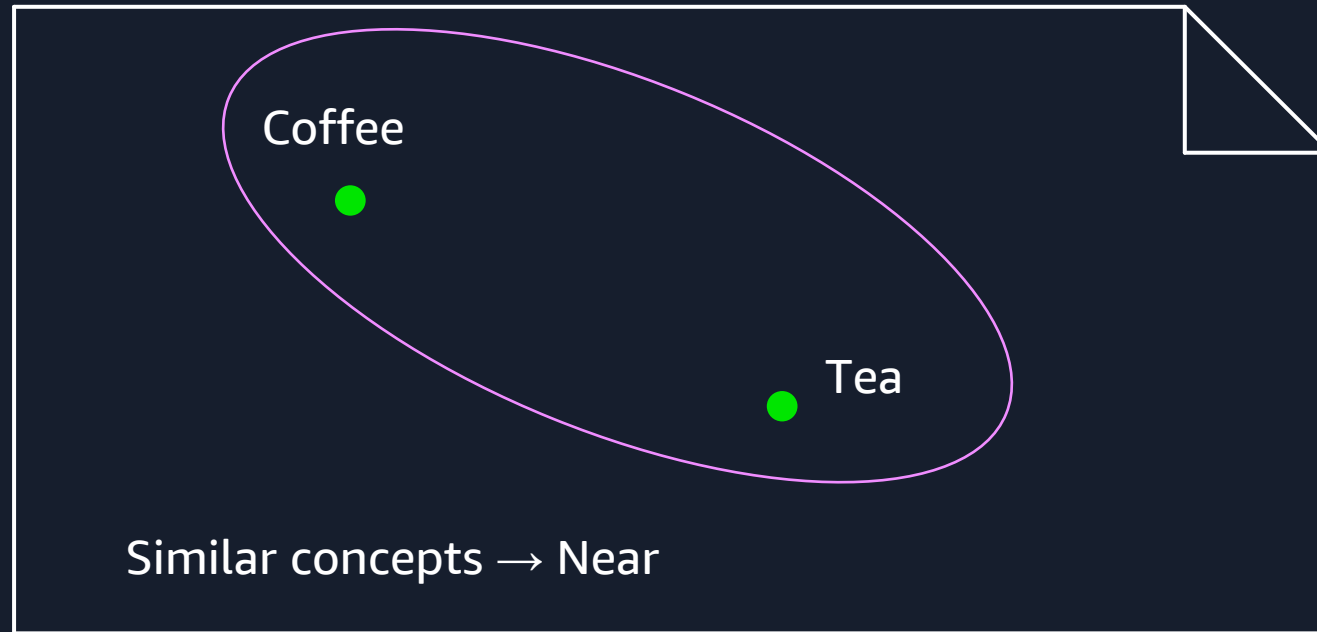
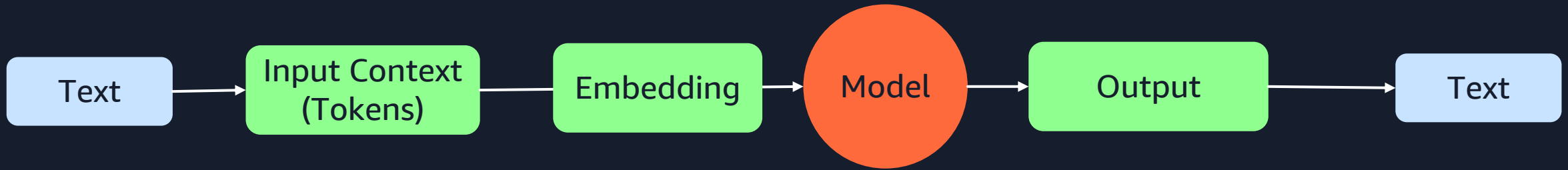


How do large language models (LLMs) work?

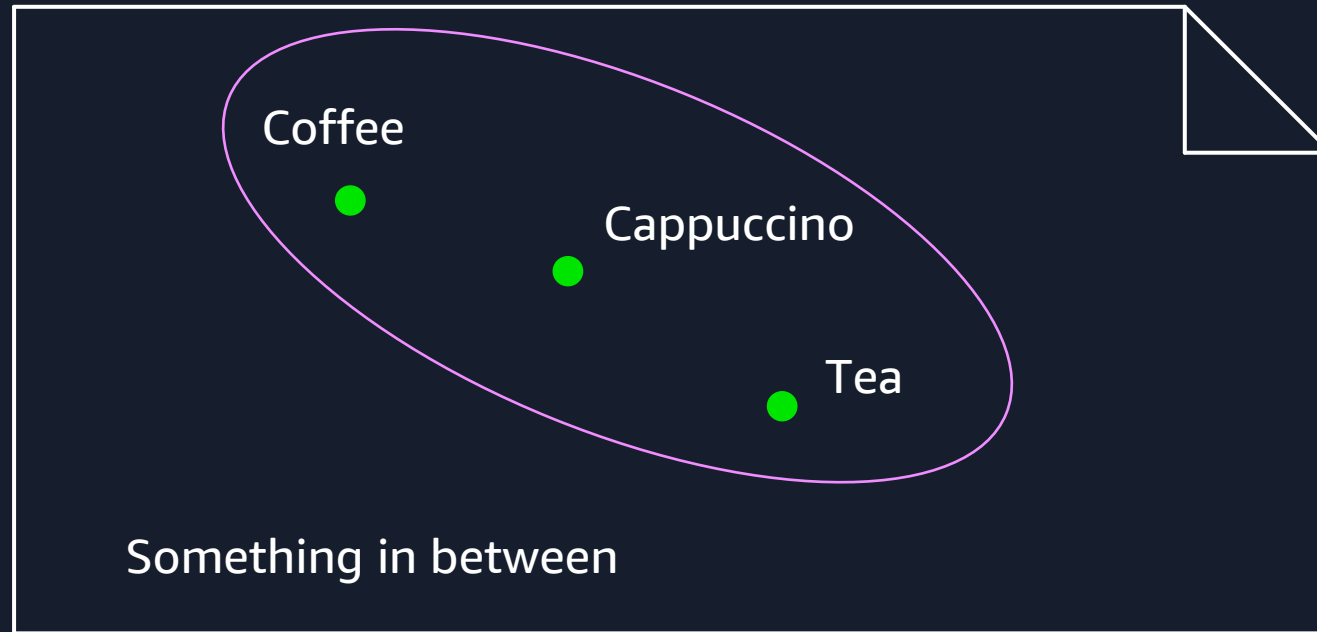
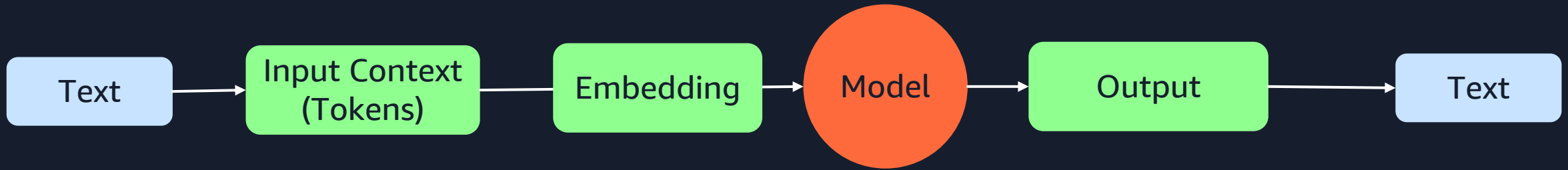
And that's why we **first tokenize** the text



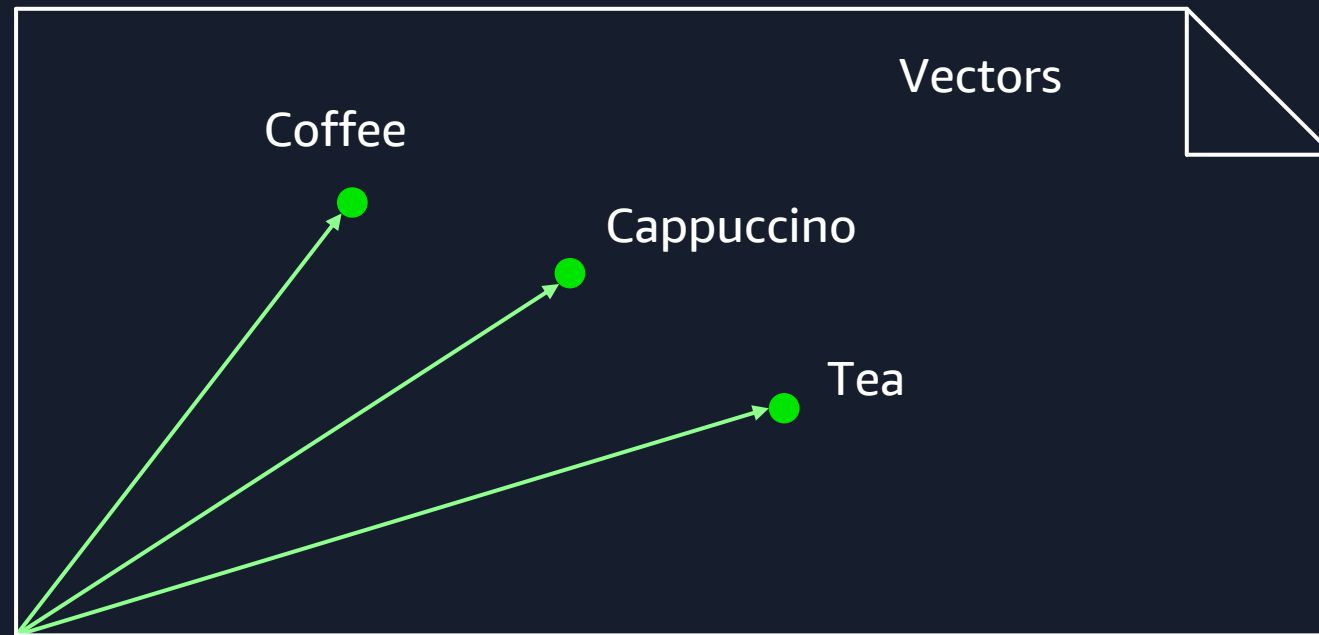
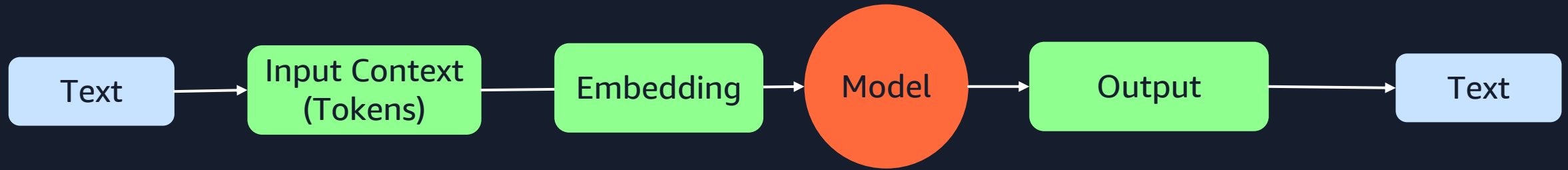
Embeddings



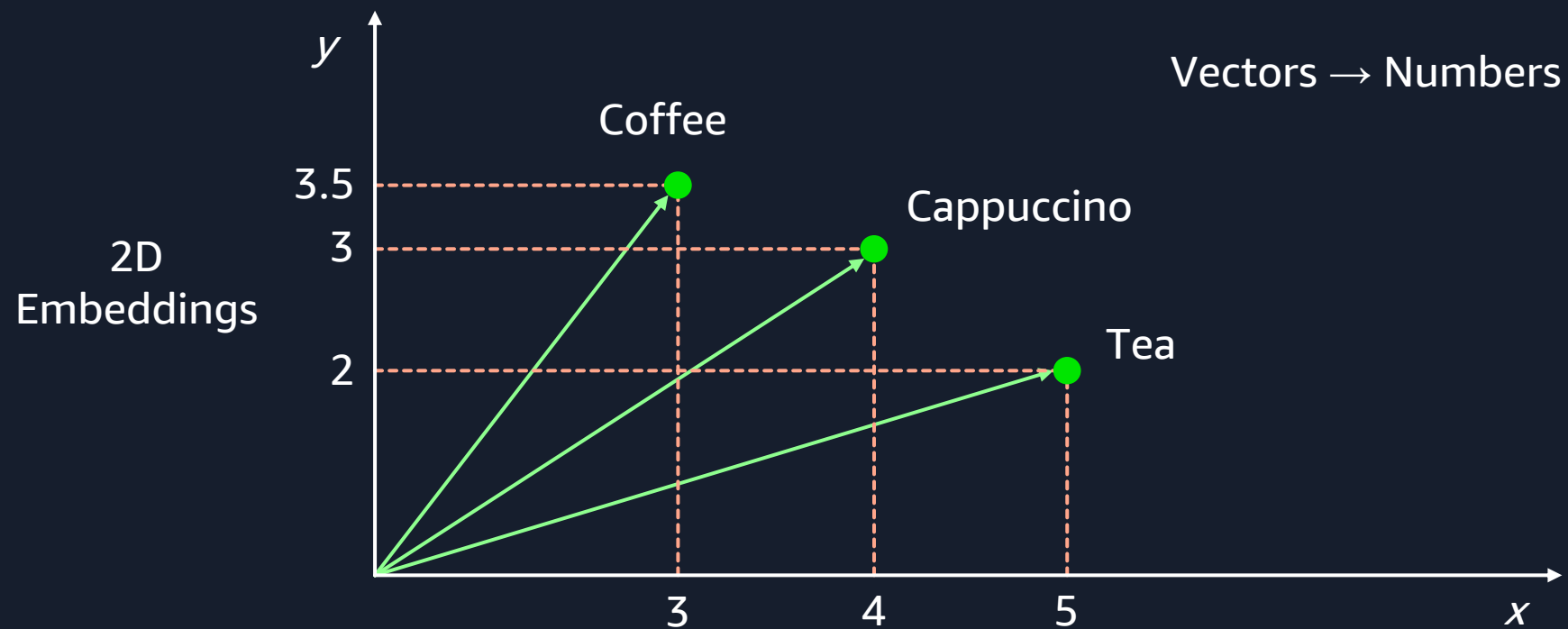
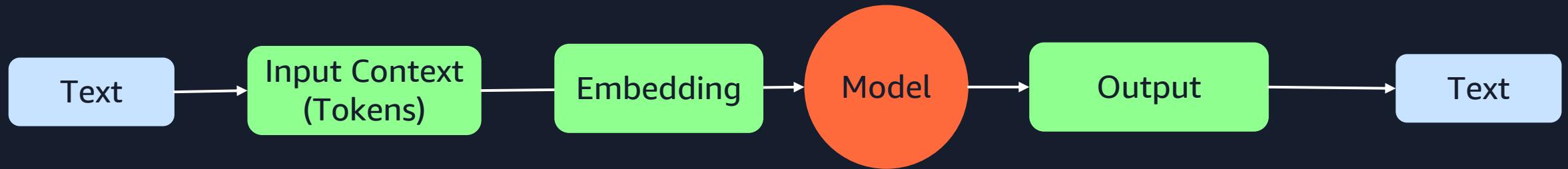
Embeddings



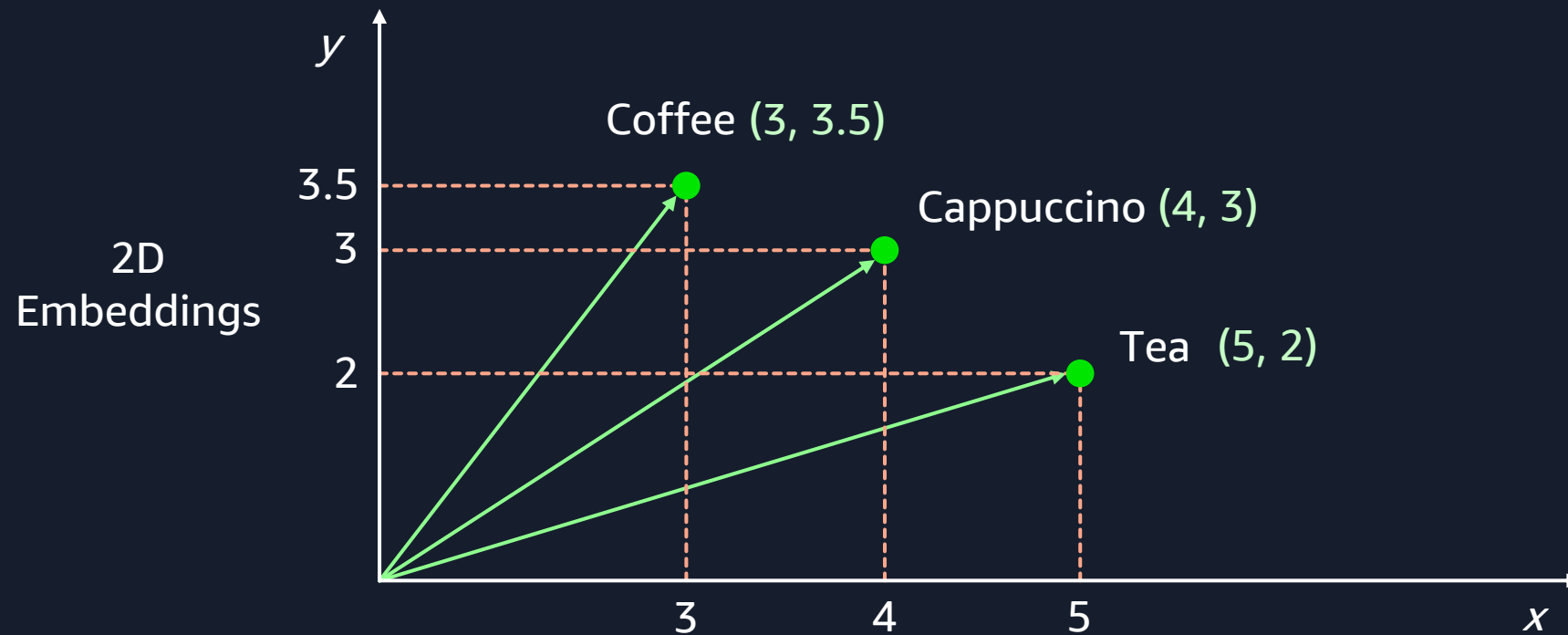
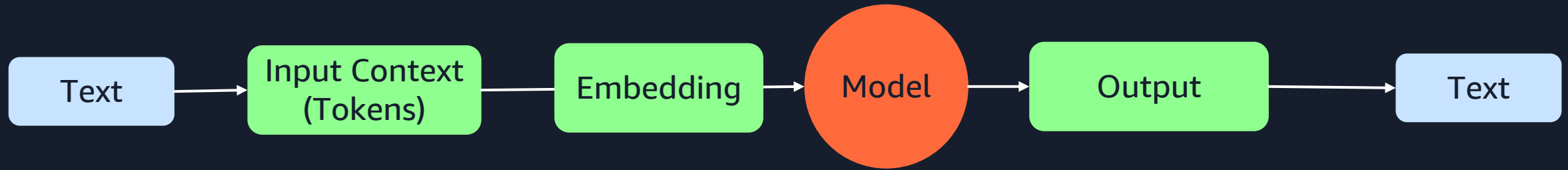
Embeddings



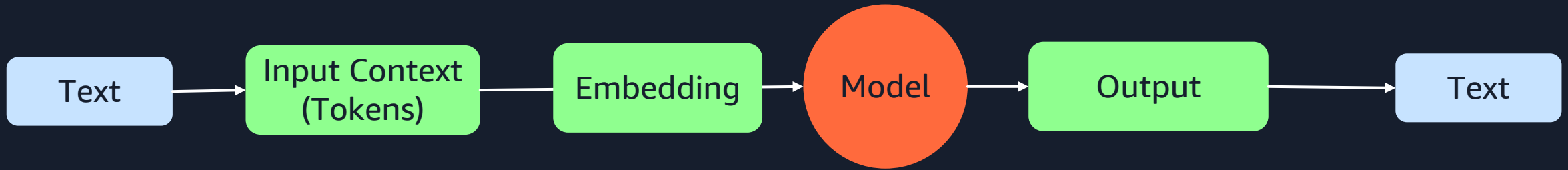
Embeddings



Embeddings



Embeddings



2D are not
enough

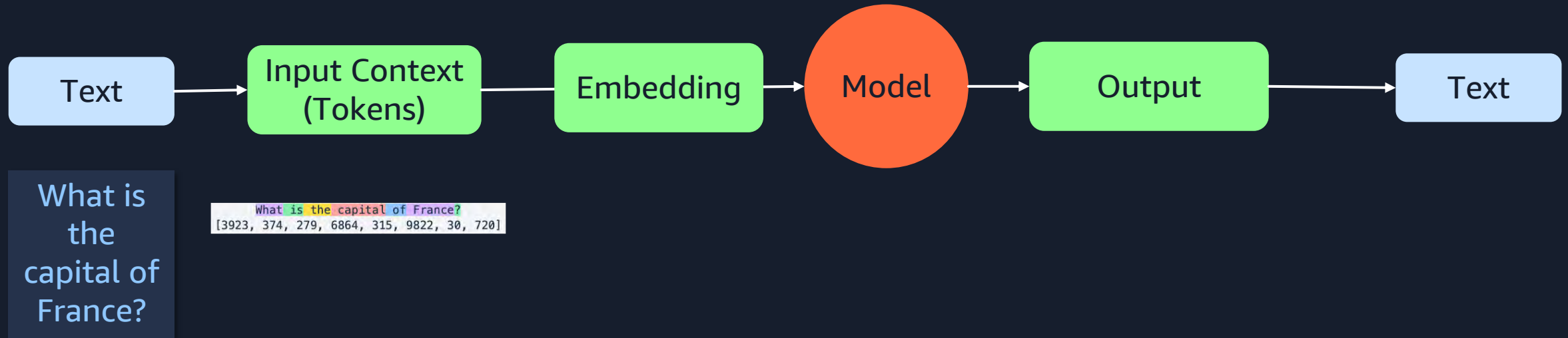
Coffee (3, 3.5, 5, 1.2, ...)

Cappuccino (4, 3, 2.7, 5, ...)

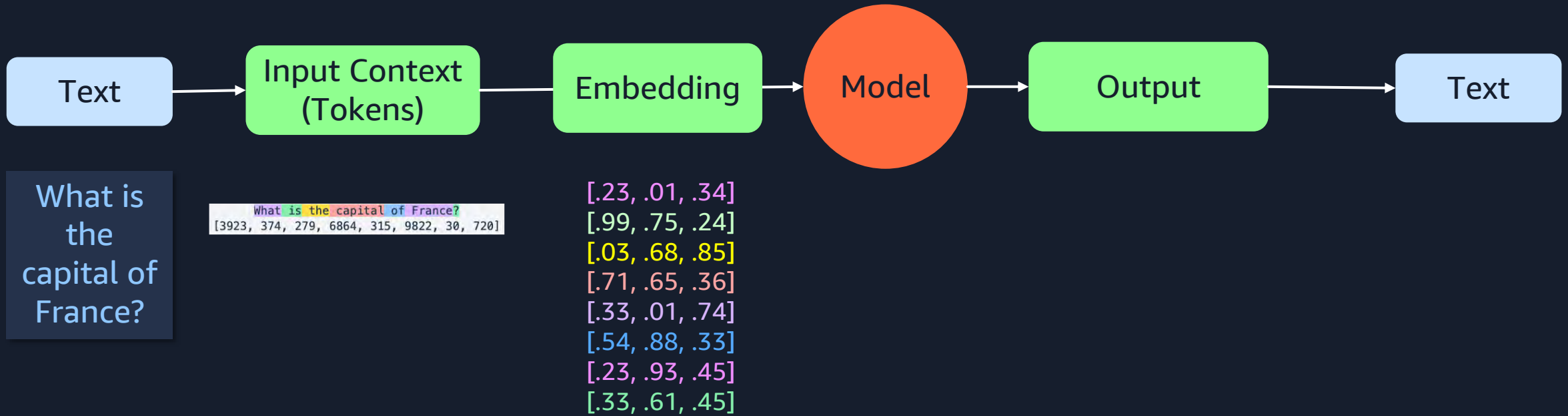
Milk (5, 2, 1.6, 4, ...)

N-Dimensional
Embeddings

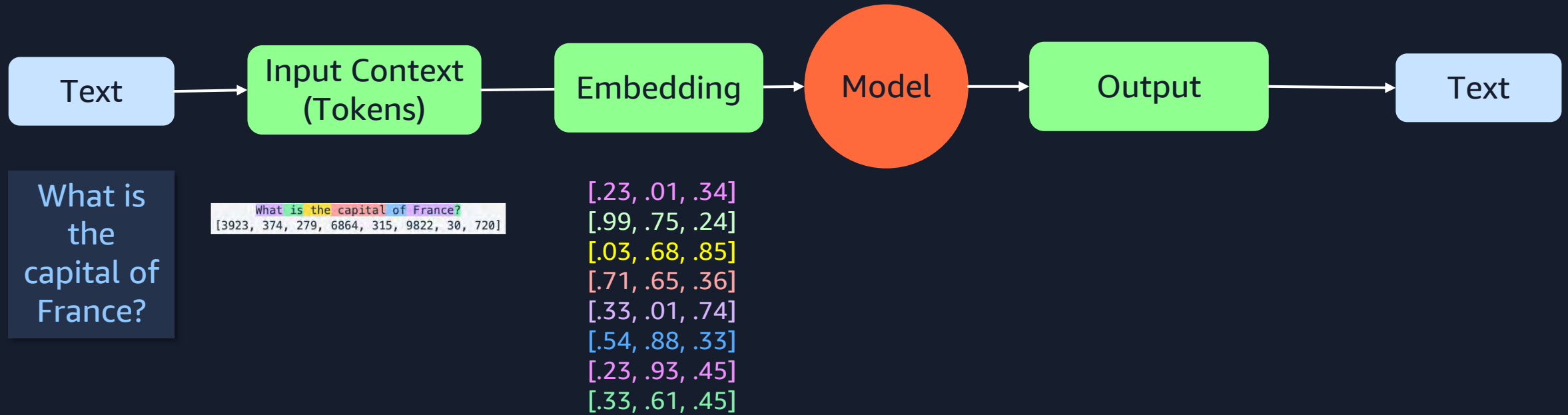
So, where we stand now ?



So, where we stand now ?

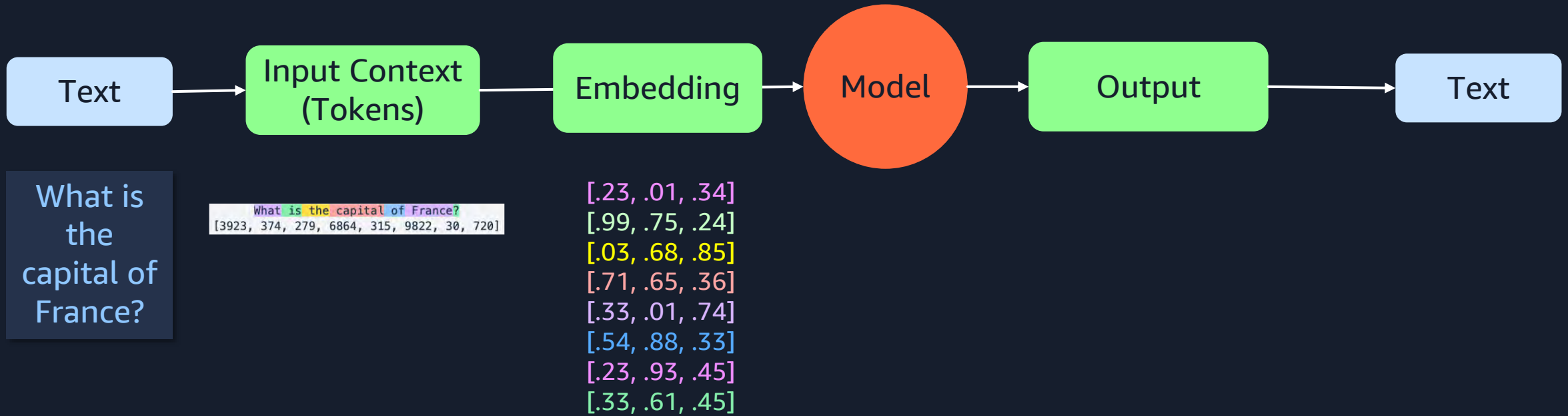


So, where we stand now ?



So, we learned that converting **our data into vectors** is the **first thing we need to do**.

So, where we stand now ?



So, we learned that converting **our data into vectors** is the **first thing we need to do**.

Now, let's think about this: can an LLM answer all of our questions ?

Vector embeddings

RAW Data

Documents



Images



Audio



Machine Learning Model
(Embedding)



Amazon Bedrock

Vector
Embedding Space

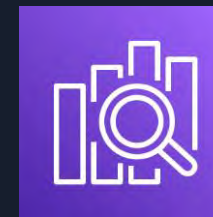
Dense Vector Encodings

0.3, 2.1, 0, 0.9, 1.0,...

1.3, 8.1, 0, 4.6, 3.6,...

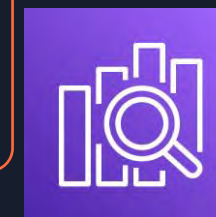
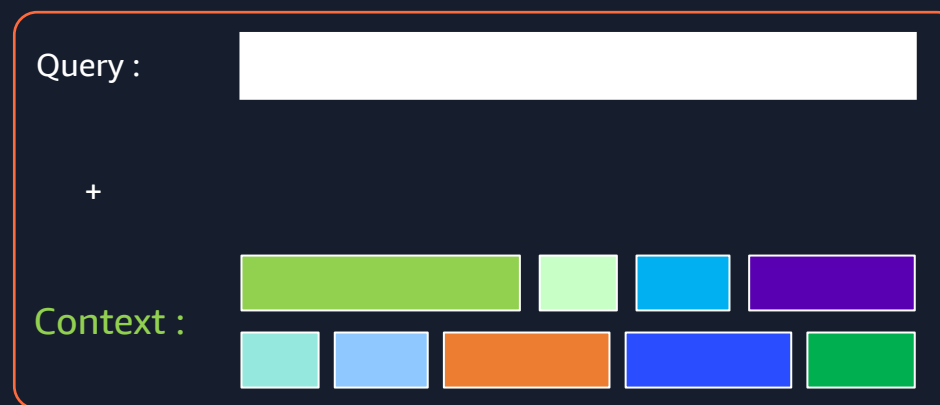
7.3, 1.1, 0, 2.9, 1.0,...

Dev ready and
Operationalized



Vector Database

Vector embeddings



Vector Database



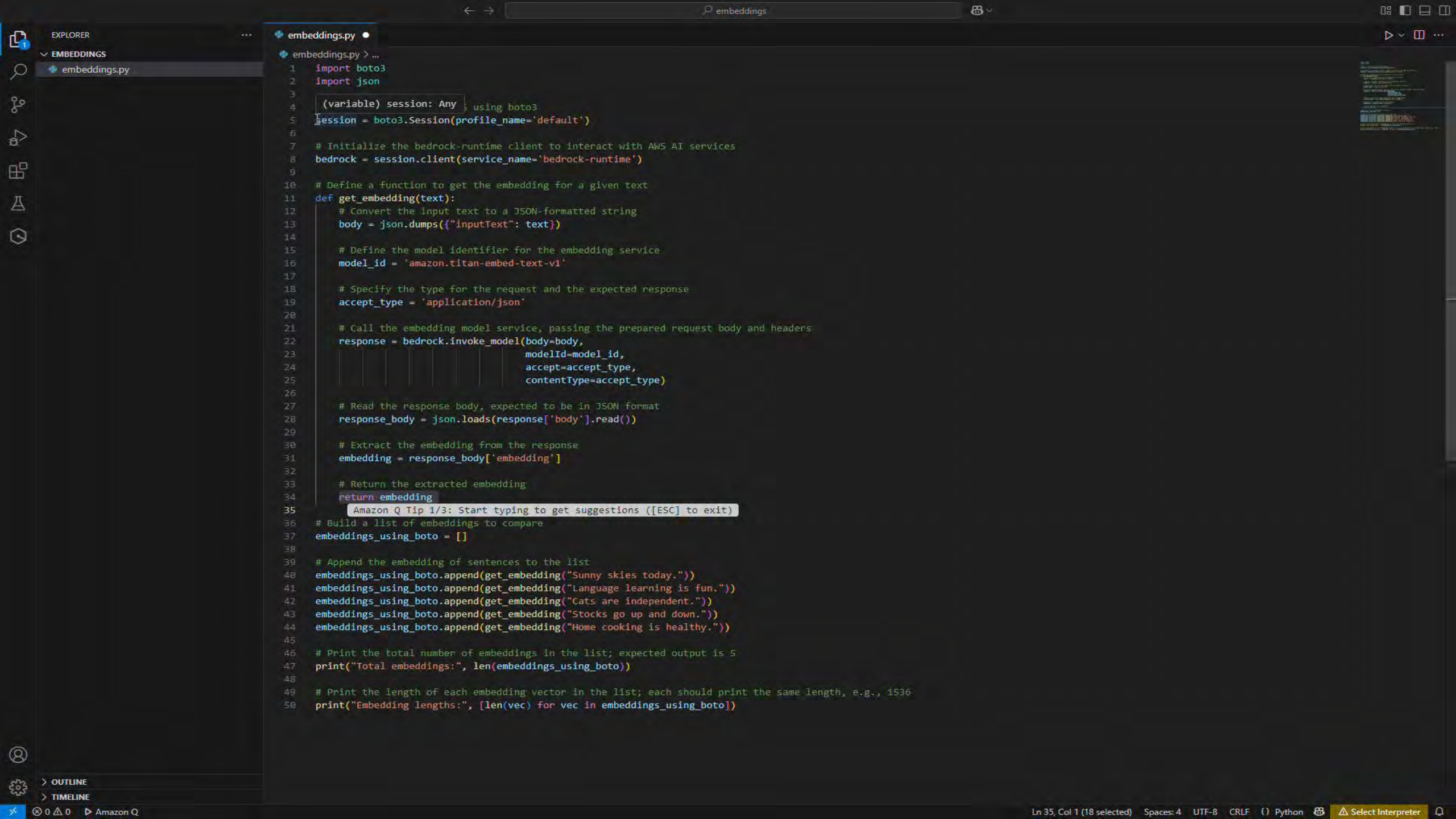
Dev ready and Operationalized

Response

Retrieval-Augmented Generation (RAG)

Demo





embeddings.py

embeddings.py > ...

```
1 import boto3
2 import json
3
4 (variable) session: Any } using boto3
5 session = boto3.Session(profile_name='default')
6
7 # Initialize the bedrock-runtime client to interact with AWS AI services
8 bedrock = session.client(service_name='bedrock-runtime')
9
10 # Define a function to get the embedding for a given text
11 def get_embedding(text):
12     # Convert the input text to a JSON-formatted string
13     body = json.dumps({"inputText": text})
14
15     # Define the model identifier for the embedding service
16     model_id = 'amazon.titan-embed-text-v1'
17
18     # Specify the type for the request and the expected response
19     accept_type = 'application/json'
20
21     # Call the embedding model service, passing the prepared request body and headers
22     response = bedrock.invoke_model(body=body,
23                                     modelId=model_id,
24                                     accept=accept_type,
25                                     contentType=accept_type)
26
27     # Read the response body, expected to be in JSON format
28     response_body = json.loads(response['body'].read())
29
30     # Extract the embedding from the response
31     embedding = response_body['embedding']
32
33     # Return the extracted embedding
34     return embedding
35
36 # Build a list of embeddings to compare
37 embeddings_using_boto = []
38
39 # Append the embedding of sentences to the list
40 embeddings_using_boto.append(get_embedding("Sunny skies today."))
41 embeddings_using_boto.append(get_embedding("Language learning is fun."))
42 embeddings_using_boto.append(get_embedding("Cats are independent."))
43 embeddings_using_boto.append(get_embedding("Stocks go up and down."))
44 embeddings_using_boto.append(get_embedding("Home cooking is healthy."))
45
46 # Print the total number of embeddings in the list; expected output is 5
47 print("Total embeddings:", len(embeddings_using_boto))
48
49 # Print the length of each embedding vector in the list; each should print the same length, e.g., 1536
50 print("Embedding lengths:", [len(vec) for vec in embeddings_using_boto])
```

Amazon Q Tip 1/3: Start typing to get suggestions ([ESC] to exit)

> OUTLINE

> TIMELINE

0 Amazon Q

Ln 35, Col 1 (18 selected) Spaces: 4 UTF-8 CRLF Python Select Interpreter

How do large language models (LLMs) work?

But do you feel, one number for each token is enough ?

Imagine, you just know the courses that are available for your internal staff at your organization

What is the capital of France?
[3923, 374, 279, 6864, 315, 9822, 30, 720]

How do large language models (LLMs) work?

But do you feel, one number for each token is enough ?

Imagine, you just know the courses that are available for your internal staff at your organization

Can anyone answer any questions related to these internal courses?

What is the capital of France?
[3923, 374, 279, 6864, 315, 9822, 30, 720]

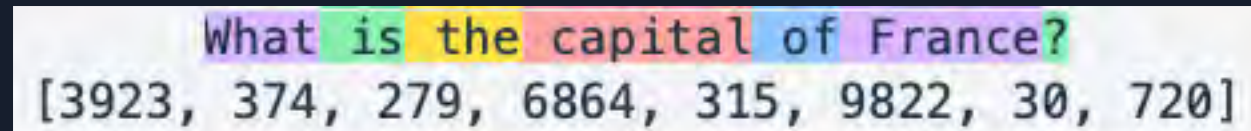
How do large language models (LLMs) work?

But do you feel, one number for each token is enough ?

Imagine, you just know that courses that are available for your internal staff at your organization

Can anyone answer any questions related to these internal courses?

How many members of staff have completed a particular course ?



What is the capital of France?
[3923, 374, 279, 6864, 315, 9822, 30, 720]

How do large language models (LLMs) work?

But do you feel, one number for each token is enough ?

Imagine, you just know that courses that are available for your internal staff at your organization

Can anyone answer any questions related to these internal courses?

How many members of staff have completed a particular course ?

How long on average does it take an employee to complete a course ?

What is the capital of France?
[3923, 374, 279, 6864, 315, 9822, 30, 720]

How do large language models (LLMs) work?

But do you feel, one number for each token is enough ?

Imagine, you just know that courses that are available for your internal staff at your organization

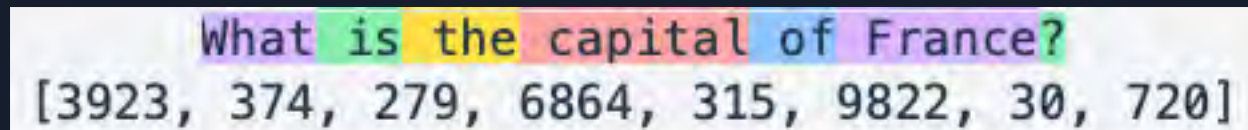
Can anyone answer any questions related to these internal courses?

How many members of staff have completed a particular course ?

How long on average does it take an employee to complete a course ?

Which course is most popular?

...

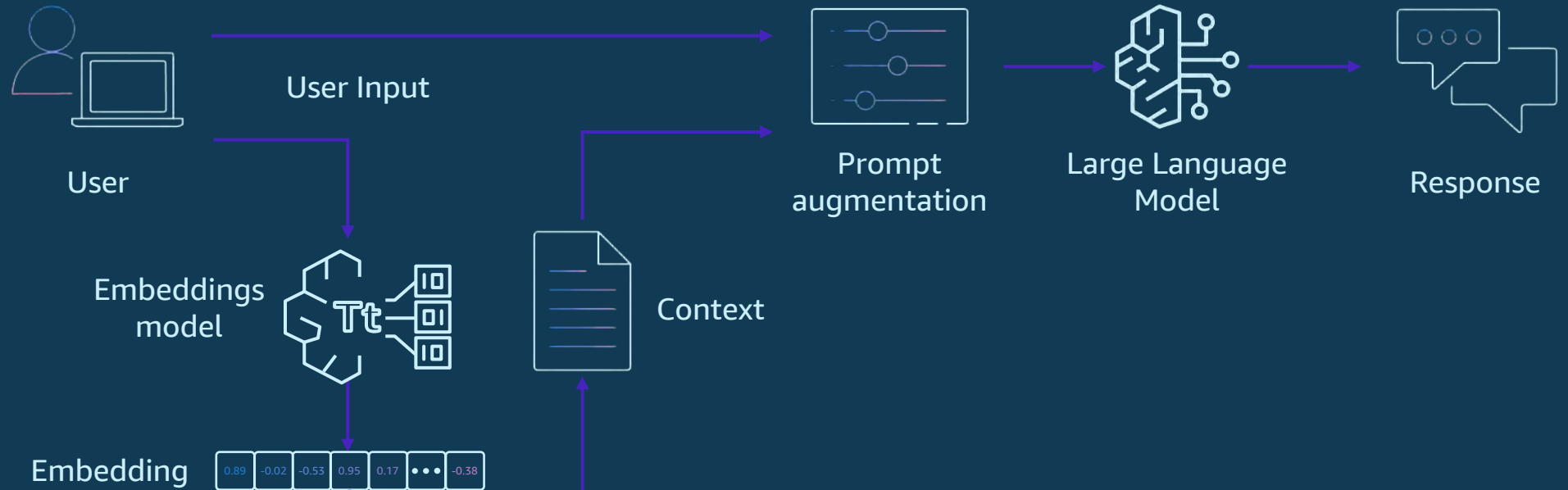


What is the capital of France?
[3923, 374, 279, 6864, 315, 9822, 30, 720]

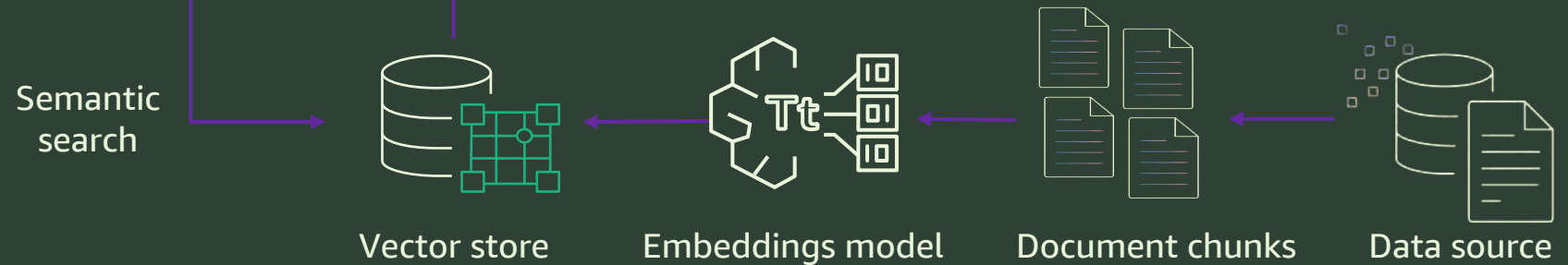


Retrieval Augmented Generation (RAG)

Text Generation Workflow

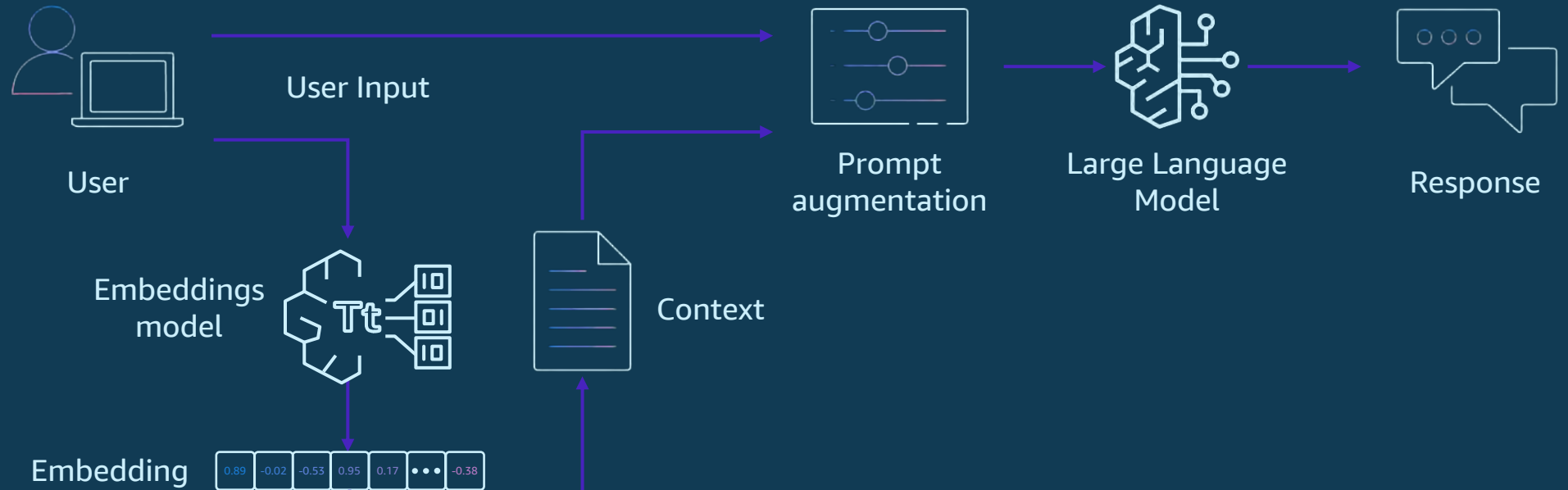


Data Ingestion Workflow

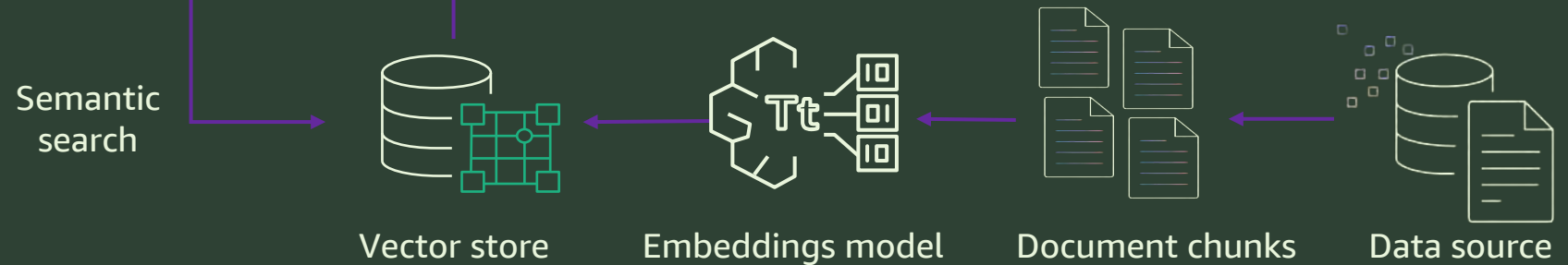


Retrieval Augmented Generation (RAG)

Text Generation Workflow



Data Ingestion Workflow



Ok, so you might feel its **too much** of task...



Managing multiple
data sources



Creating vector embeddings
for large volumes of data



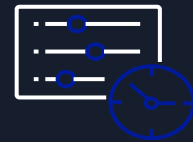
Incremental updates
to vector store



Coding effort



Scaling retrieval mechanism



Orchestration

Amazon Bedrock Knowledge Bases

Gives FMs and agents **contextual information** from your private data sources for RAG



Fully managed support for end-to-end RAG workflow



Securely connect FMs and agents to data sources

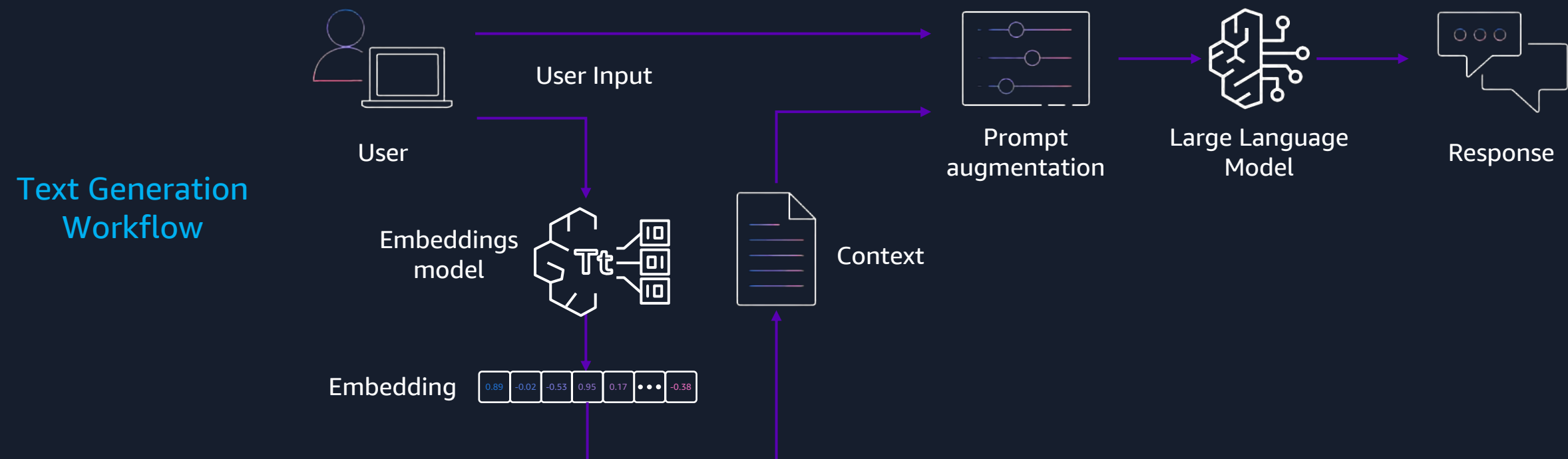


Easily retrieve relevant data and augment prompts



Provide source attribution

Amazon Bedrock Knowledge Bases

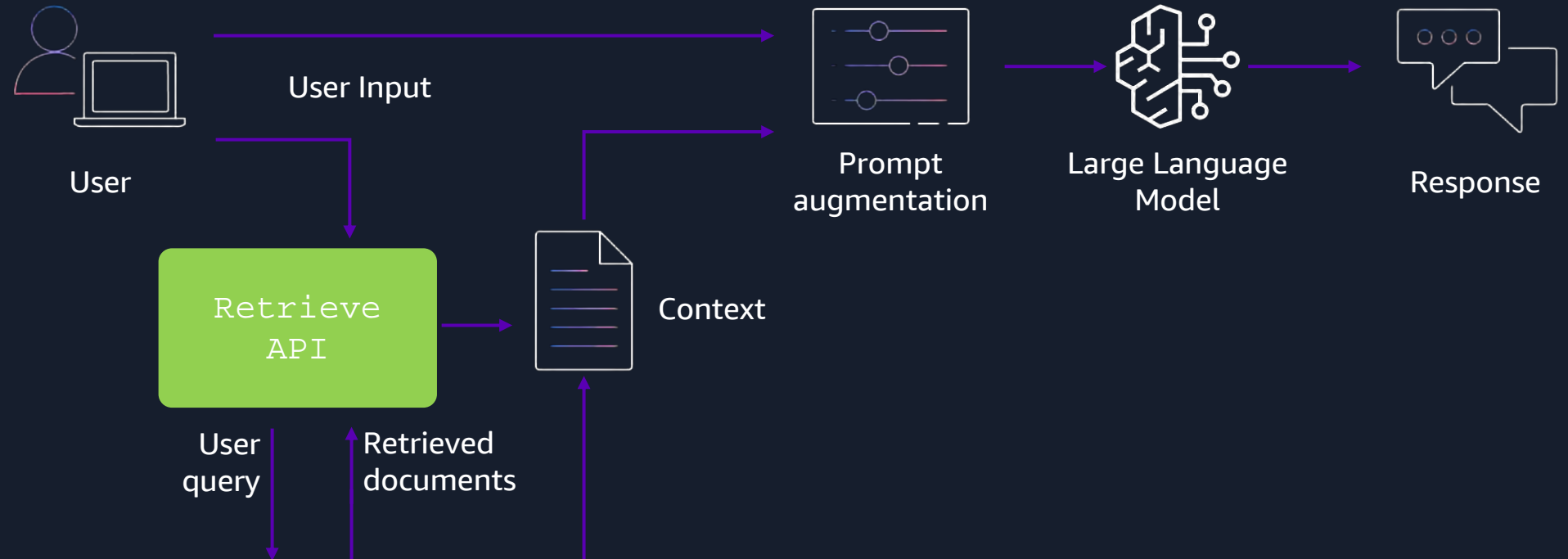


Amazon Bedrock Knowledge Bases



Amazon Bedrock Knowledge Bases (Retrieve API)

Text Generation Workflow



Data Ingestion Workflow

Amazon Bedrock Knowledge Bases

Vector store

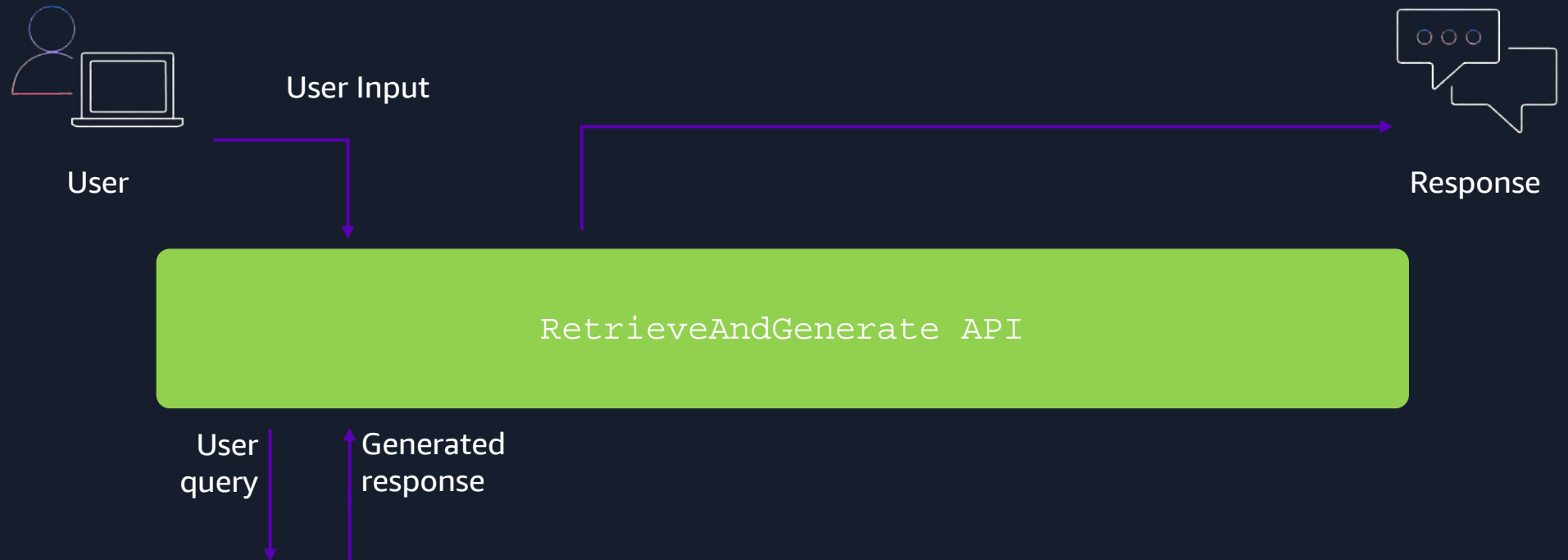
Embeddings model

Document chunks

Data source

Amazon Bedrock Knowledge Bases (RetrieveAndGenerate API)

Fully managed
RAG



Amazon Bedrock Knowledge Bases



Get started with
Amazon Bedrock



What are Large
Language Models



Multimodal RAG & Embeddings
with
Amazon Nova & Bedrock

Thank you!

Veliswa Boya

Senior Developer Advocate, AWS

