Pierian Training

Course Curriculum

Gemini Python API

- Welcome to this free course, which will be your quick start guide to using the Python API for the latest Gemini AI model from Google!
- Please keep in mind a few things:
    - Prerequisites on the landing page (Experience requirements and GMail).
    - Our 2 hour time limit.

Gemini Python API

- **Course Curriculum**
  - Understanding LLMs and API Access
  - Text Generation
    - Text Generation
    - Chat Models
    - Configuration Parameters
  - Vision Model and Multimodal Inputs
  - RAG - Retrieval Augmented Generation

Let’s get started!

How an LLM Works

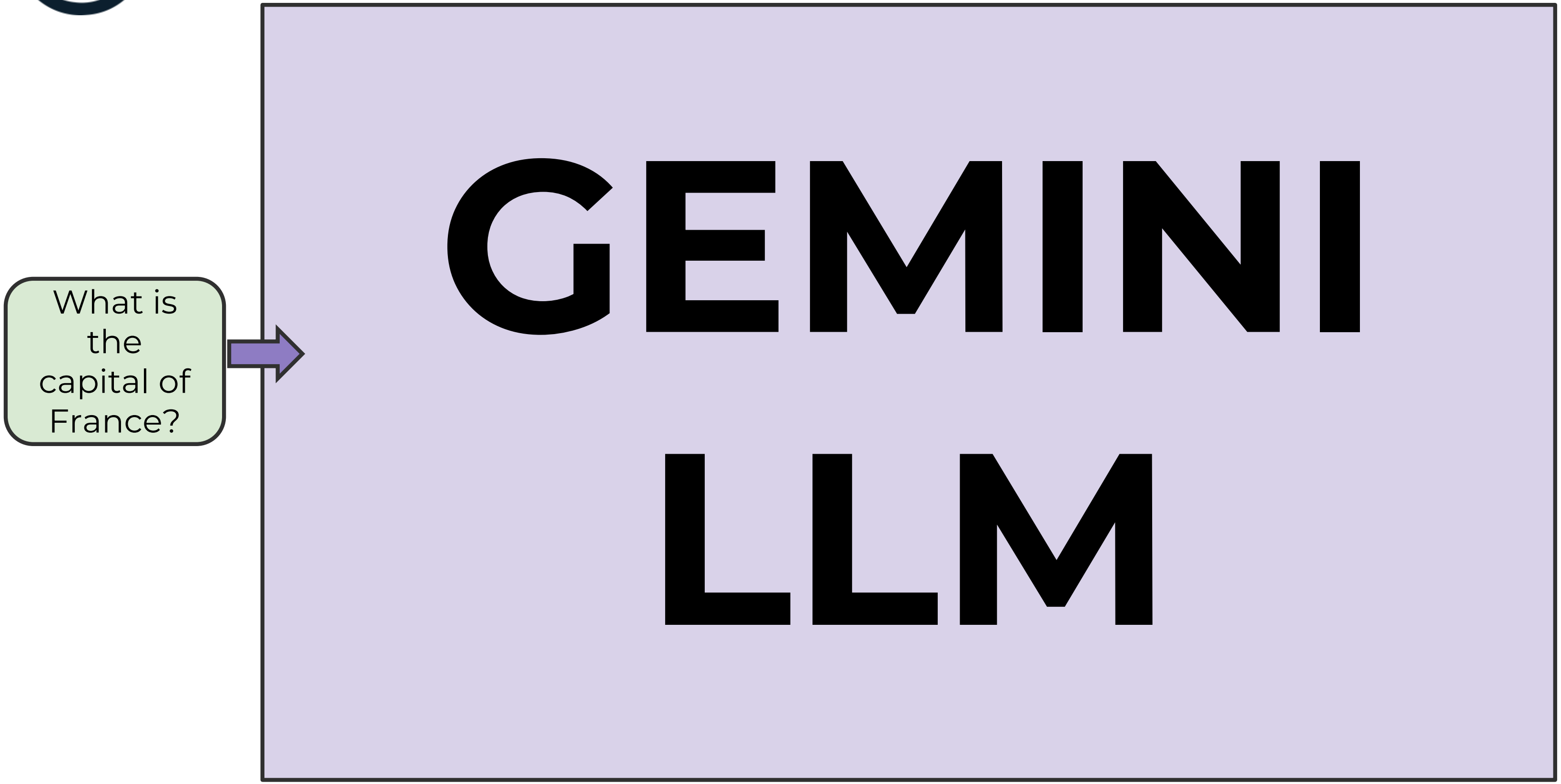
Gemini Python API

- Let’s go over how an LLM works from a very high level overview, note that we are discussing **inference** in this lecture, not **training**.
- Having a basic understanding of how the model works will help you understand the generation configuration parameters we’ll cover later on in the course!

Gemini Python API

What is the capital of France?

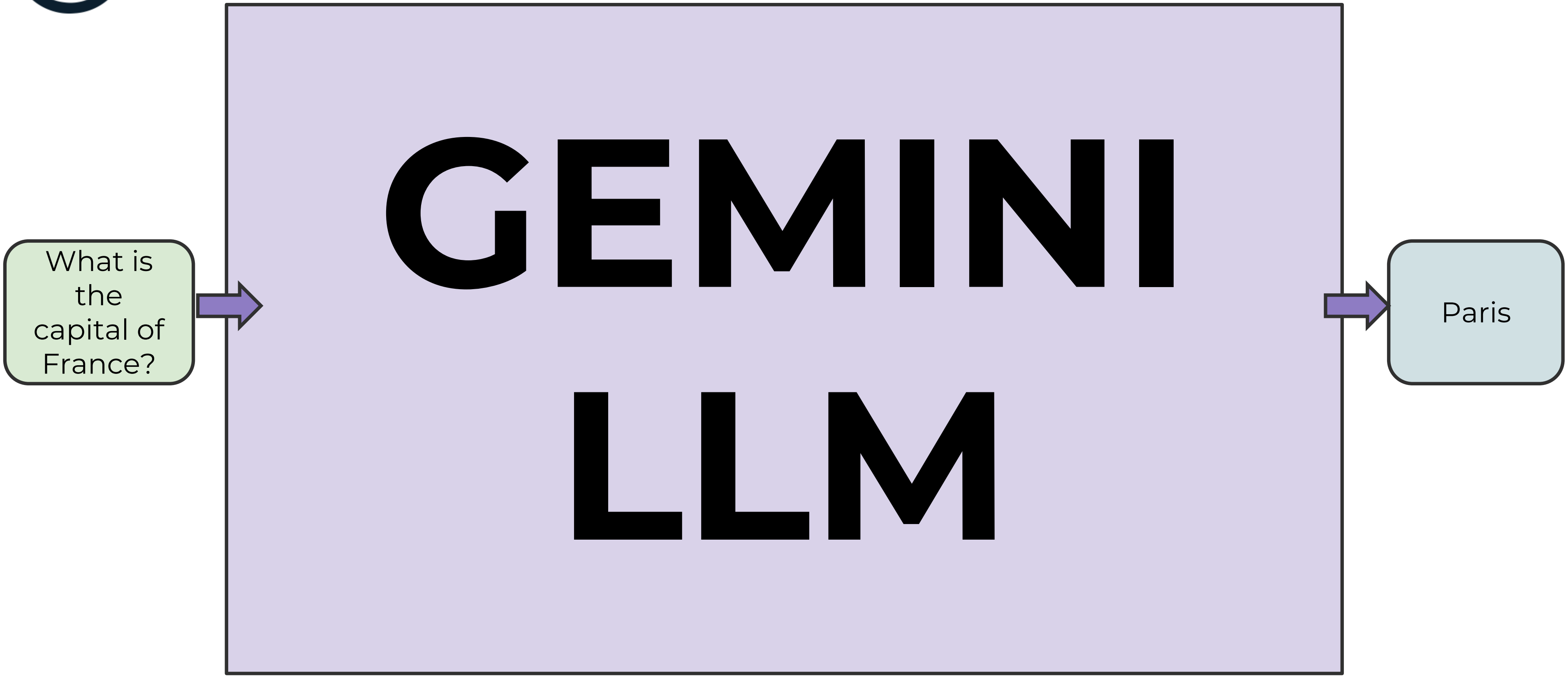
Gemini Python API



What is the capital of France?

GEMINI LLM

Gemini Python API



What is the capital of France?

GEMINI LLM

Paris

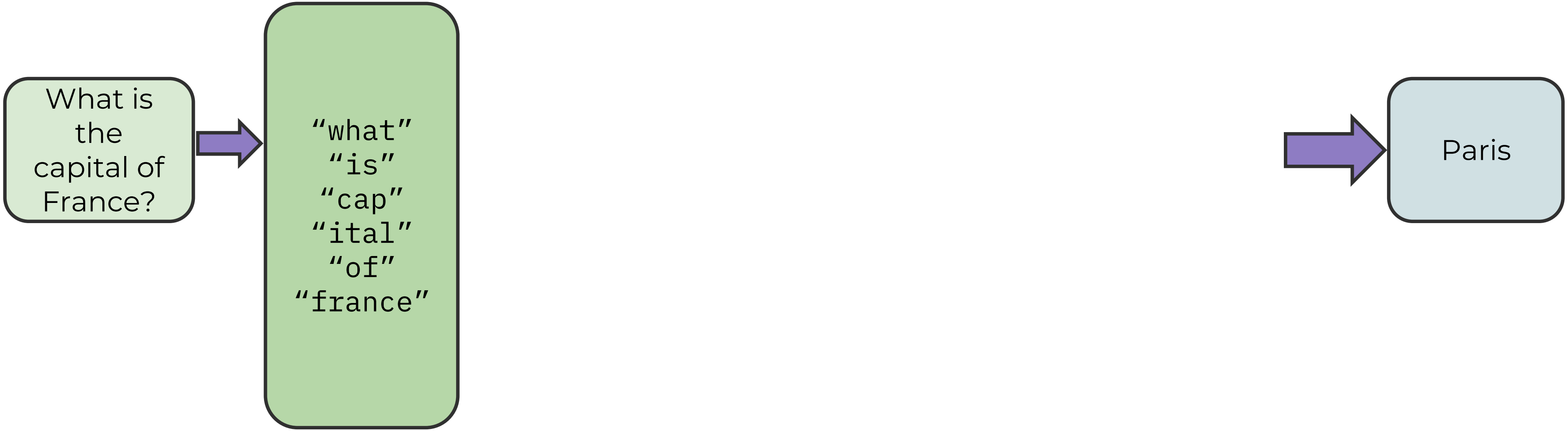
Gemini Python API

What is the capital of France?

Paris

Gemini Python API

Tokenization

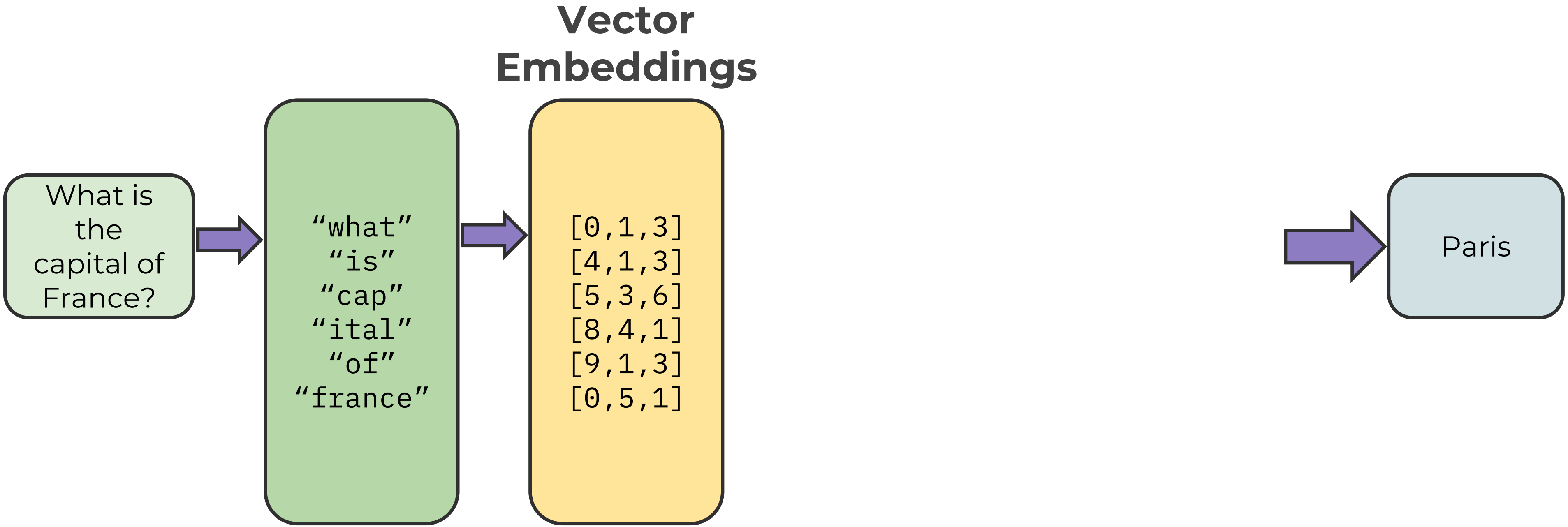


What is the capital of France?

“what”
“is”
“cap”
“ital”
“of”
“france”

Paris

Gemini Python API



Vector Embeddings

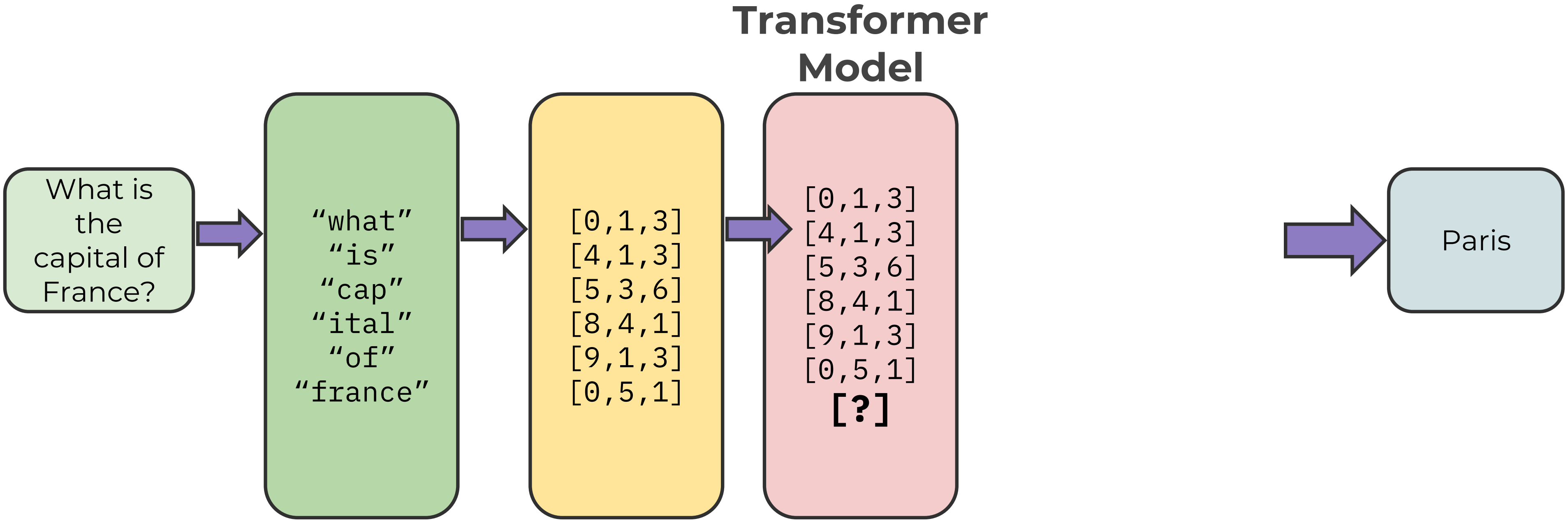
What is the capital of France?

“what”
“is”
“cap”
“ital”
“of”
“france”

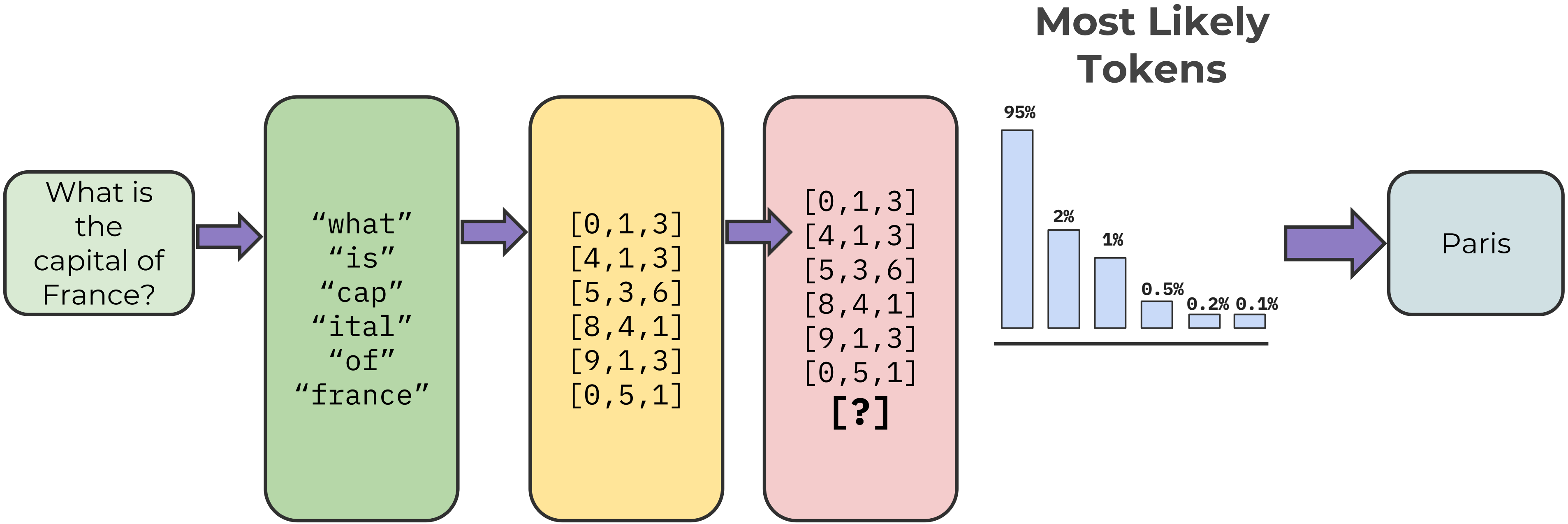
[0,1,3]
[4,1,3]
[5,3,6]
[8,4,1]
[9,1,3]
[0,5,1]

Paris

Gemini Python API



Gemini Python API



What is the capital of France?

```
"what"
"is"
"cap"
"ital"
"of"
"france"
```

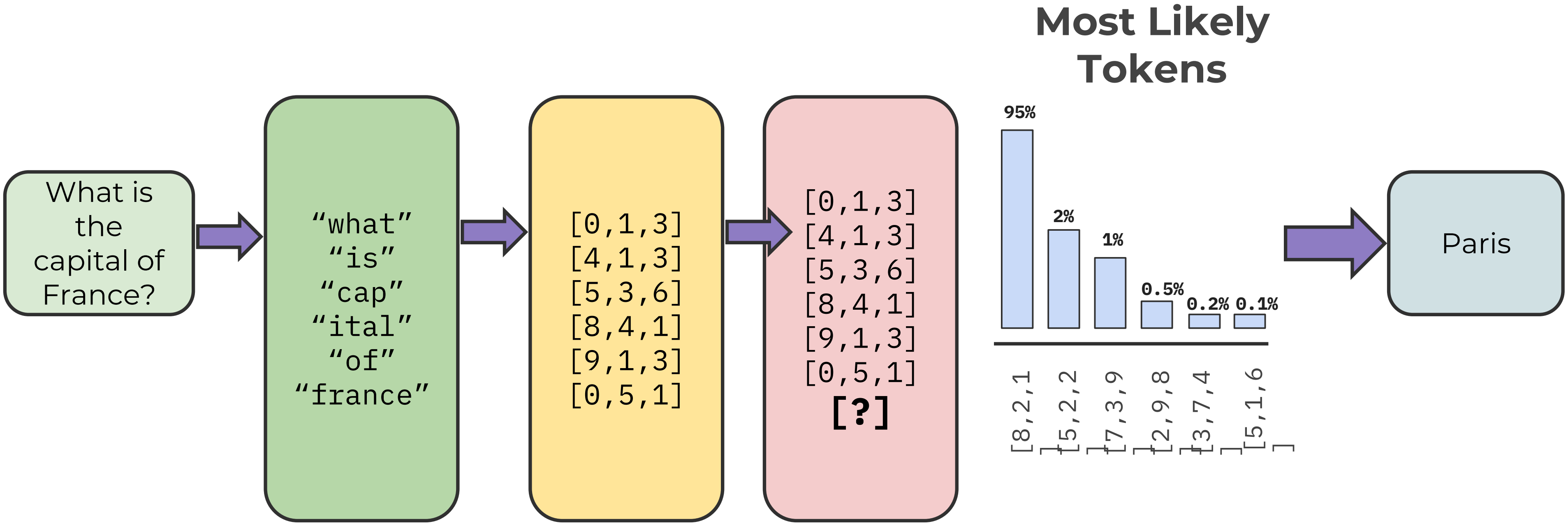
```
[0,1,3]
[4,1,3]
[5,3,6]
[8,4,1]
[9,1,3]
[0,5,1]
```

```
[0,1,3]
[4,1,3]
[5,3,6]
[8,4,1]
[9,1,3]
[0,5,1]
[?]
```

Most Likely Tokens

Paris

Gemini Python API



What is the capital of France?

“what”
“is”
“cap”
“ital”
“of”
“france”

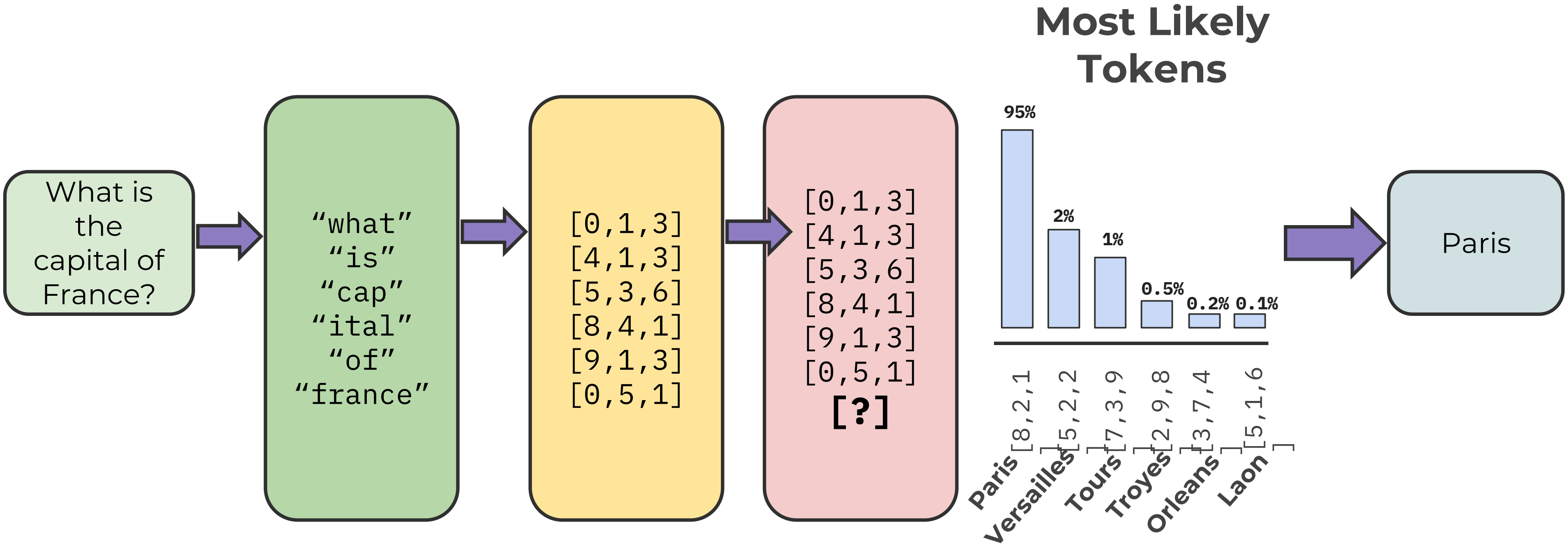
```
[0,1,3]
[4,1,3]
[5,3,6]
[8,4,1]
[9,1,3]
[0,5,1]
```

```
[0,1,3]
[4,1,3]
[5,3,6]
[8,4,1]
[9,1,3]
[0,5,1]
[?]
```

Most Likely Tokens

Paris

Gemini Python API



What is the capital of France?

“what”
“is”
“cap”
“ital”
“of”
“france”

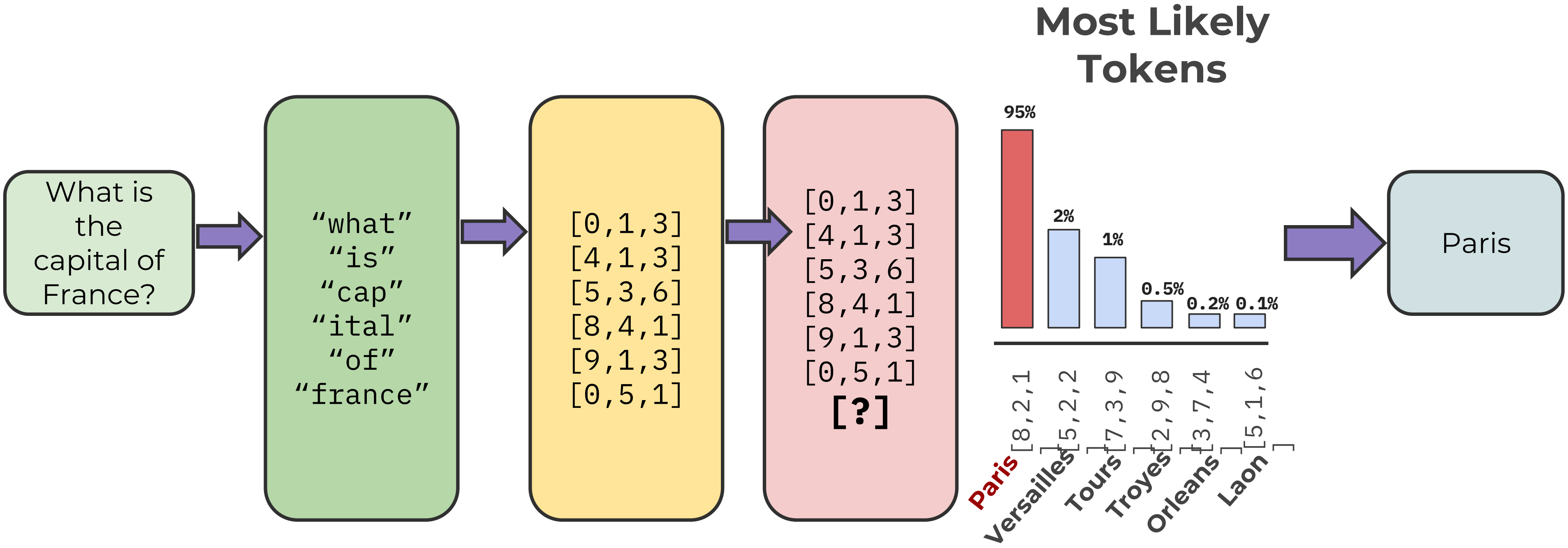
```
[0,1,3]
[4,1,3]
[5,3,6]
[8,4,1]
[9,1,3]
[0,5,1]
```

```
[0,1,3]
[4,1,3]
[5,3,6]
[8,4,1]
[9,1,3]
[0,5,1]
[?]
```

Most Likely Tokens

Paris

Gemini Python API



What is the capital of France?

“what”
“is”
“cap”
“ital”
“of”
“france”

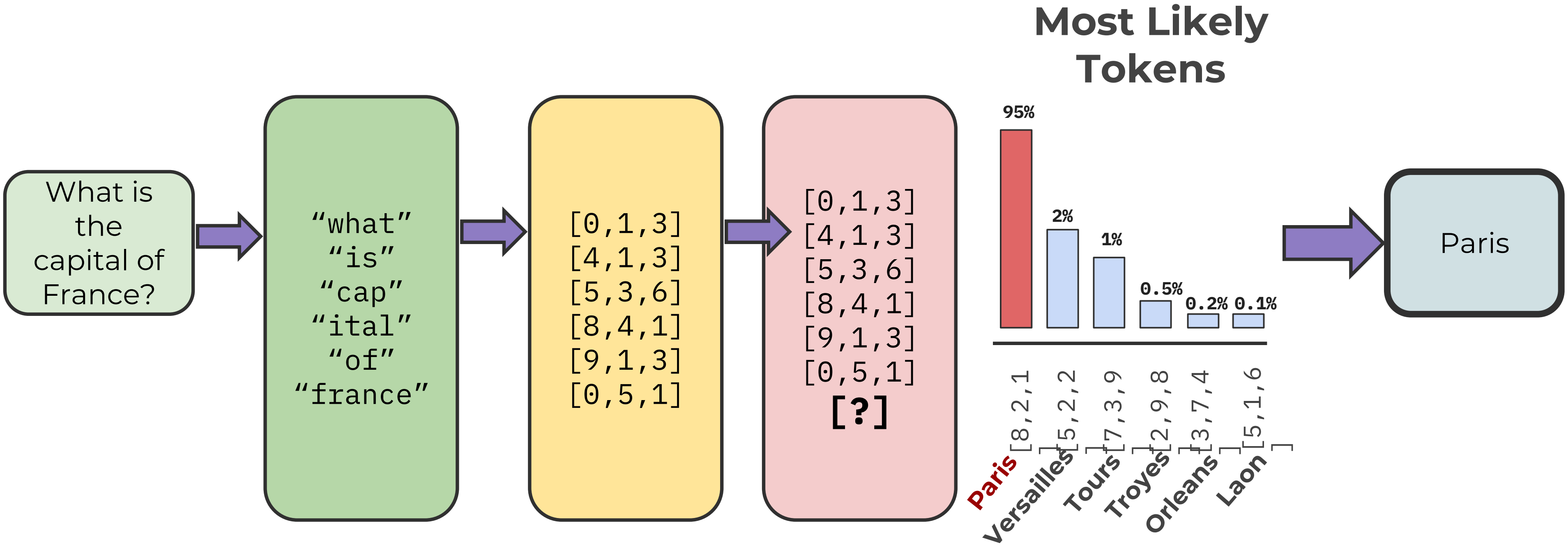
[0,1,3]
[4,1,3]
[5,3,6]
[8,4,1]
[9,1,3]
[0,5,1]

[0,1,3]
[4,1,3]
[5,3,6]
[8,4,1]
[9,1,3]
[0,5,1]
[?]

Most Likely Tokens

Paris

Gemini Python API



What is the capital of France?

“what”
“is”
“cap”
“ital”
“of”
“france”

[0,1,3]
[4,1,3]
[5,3,6]
[8,4,1]
[9,1,3]
[0,5,1]

[0,1,3]
[4,1,3]
[5,3,6]
[8,4,1]
[9,1,3]
[0,5,1]
[?]

Most Likely Tokens

Paris

Gemini Python API

- **Key Takeaways:**
  - Model uses tokens, not words.
  - Model has its own internal vector embedding representation of tokens.
  - The next most probable token is chosen from a distribution, allowing for stochastic results (next token is not deterministic, even for the exact same input).

Gemini Python API

- **Later on we'll explore configuration parameters and RAG with embeddings!**
  - We can take advantage of the model's ability to embed words to vectors for RAG - Retrieval Augmented Generation.
  - We can edit configuration parameters, for example, changing how we create the probability distribution of the next most likely token.

Let’s get started!

API Access

Gemini Python API

- **API Access**
  - To begin using Python to access the latest Gemini models, we can go to:
    - **ai.google.dev**
  - There are technically two ways to access the Gemini models:
    - API Key in Google AI Studio
    - Vertex AI via Google Cloud

Gemini Python API

- **API Access via Google Cloud**
  - Google Cloud has already had hosted LLMs like the PaLM model, meaning it has more advanced IAM capabilities for access management.
  - We won’t cover this approach in this course, but you can learn more at:
    - **https://cloud.google.com/vertex-ai/docs/reference/rest**

Gemini Python API

- **API Access via Google Cloud**
  - We only recommend this approach if you’ve already used Google Cloud Python SDKs, since it requires creating a Google Cloud account, creating an IAM profile, and download JSON credentials for that account with Vertex AI permissions.

Gemini Python API

- **API Key in Google AI Studio**
  - Google has an easy to use “Google AI Studio” located at:
    - **makersuite.google.com**
  - This studio contains both an API Key creation center and a graphical interface to test prompts.

Gemini Python API

- **Let’s continue by going to:**
  - **ai.google.dev**
- Since this URL may change in the future, you may want to perform a quick Google search to confirm with “Google AI Studio” or “Google Gemini API Key”.
- Make sure to check our notebook for troubleshooting links and other helpful information!