Pierian Training

Text Generation

Chat

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

Configuration Parameters

Gemini Python API

- **Configuration Parameters**
  - Gemini allows you to configure some parameters to change the output results of the model:
    - Temperature
    - Max Output Tokens
    - Top K and Top P
    - Stop Sequences
    - Candidate Count (currently only 1)

- **Max Output Tokens**
  - The amount of output tokens is set to the max by default (8192 tokens for Gemini Pro).
  - However you can try to get shorter responses or cut-off responses by setting the maximum output tokens to a lower value.

Gemini Python API

- **Stop Sequences**
  - You can specify a list of stop sequence values to stop the text generation, for example, if you are asking for a SQL query, you may set a semicolon as the stop sequence, to make sure Gemini doesn’t continue pass the query with an additional explanation.

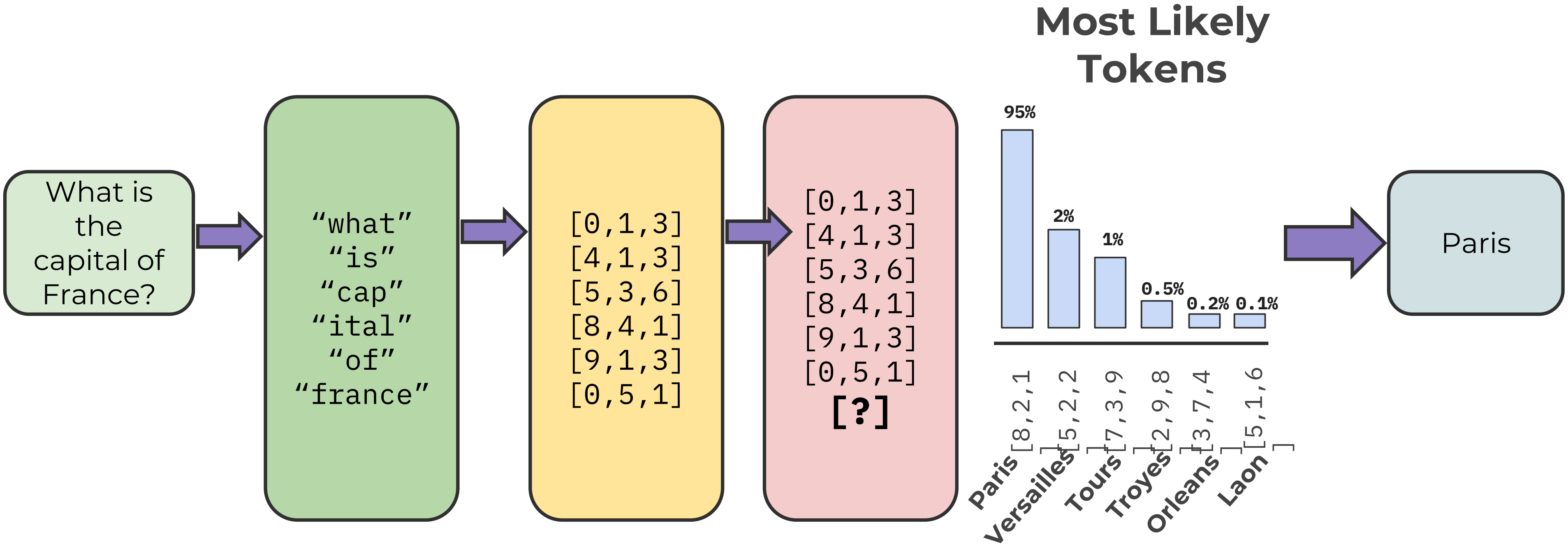
Gemini Python API

- **Candidate Count**
  - Currently, Gemini is limited to one candidate response, but in the future, the Gemini model will allow you to ask for multiple candidates to a single prompt.

Gemini Python API

- To best understand the Temperature, Top K, and Top P parameters, recall our discussion on how LLMs work, where we described the LLM creating a probability distribution

Gemini Python API



What is the capital of France?

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

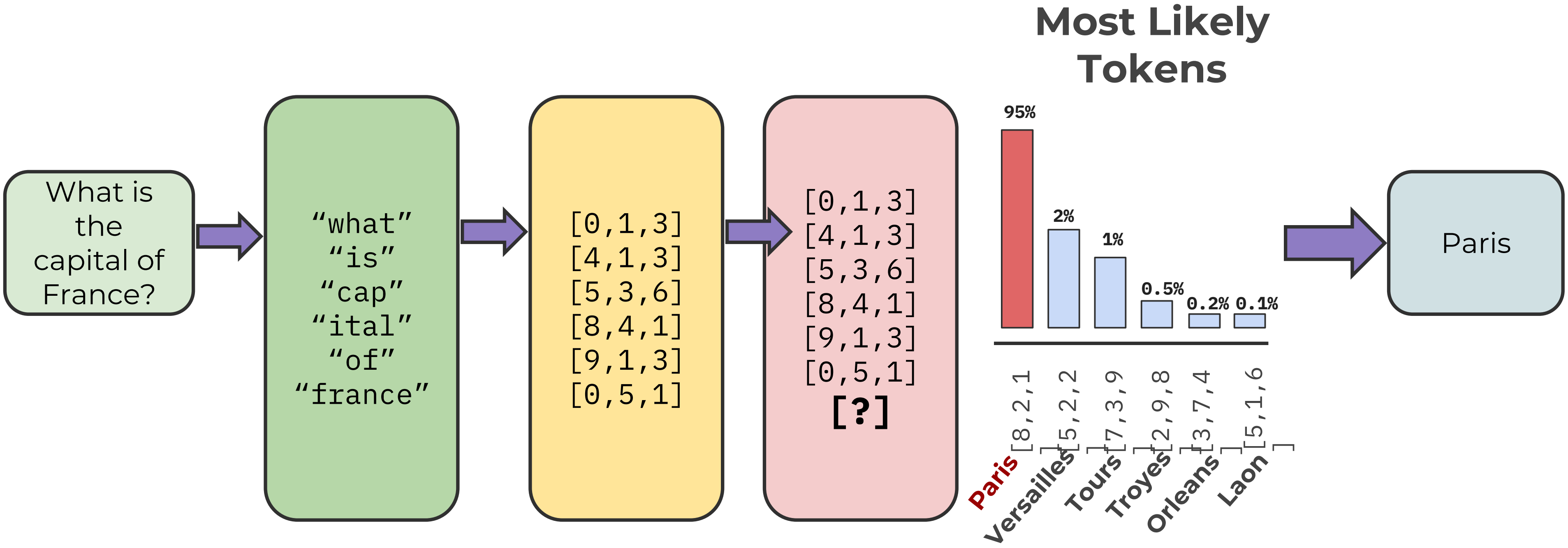
[0,1,3]
[4,1,3]
[5,3,6]
[8,4,1]
[9,1,3]
[0,5,1]

[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

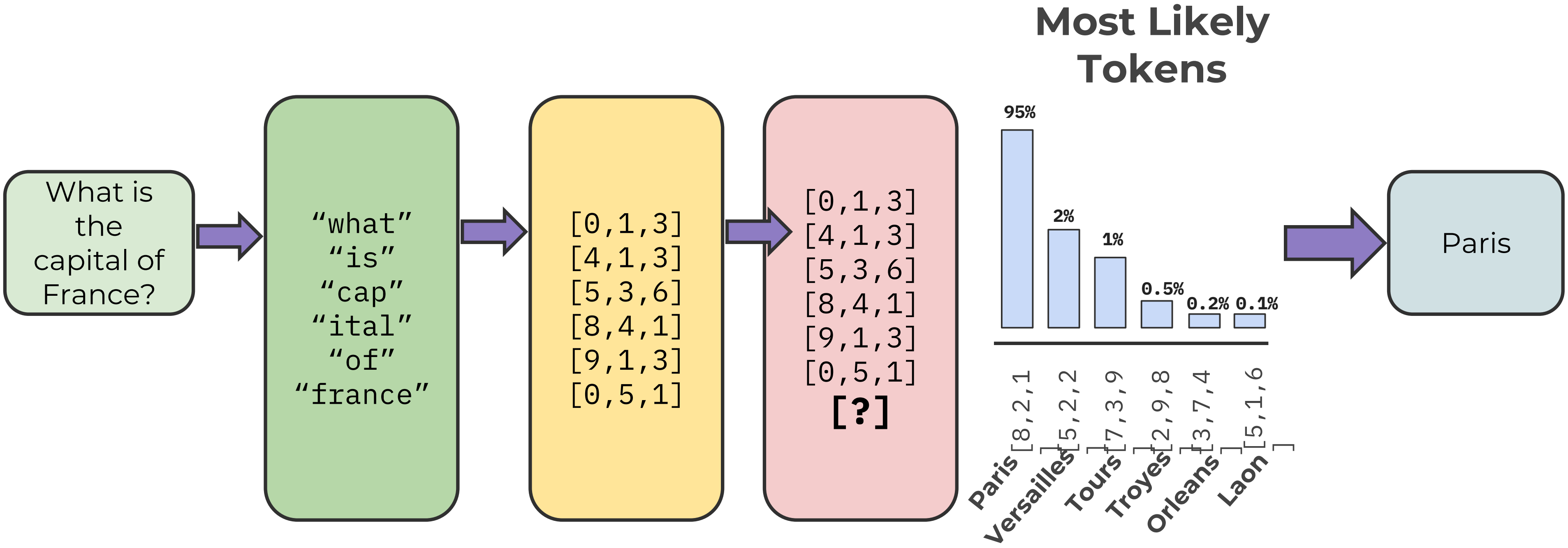
Gemini Python API

- **Temperature**
  - The term temperature comes from statistical thermodynamics.
  - You can think of this as effecting the sampling of the distribution of tokens.
  - Lower temperatures will cause the model sample the most likely tokens while a higher temperature will push the model to sample less likely tokens.

Gemini Python API

- **Temperature**
  - In other words:
    - **Higher Temperature (~1.0)**
      - More “creative” results, could sometimes go off topic or random.
    - **Lower Temperature (~0.0)**
      - Less “creative” results, should be used in situations where you expect a singular correct answer.

Gemini Python API



What is the capital of France?

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

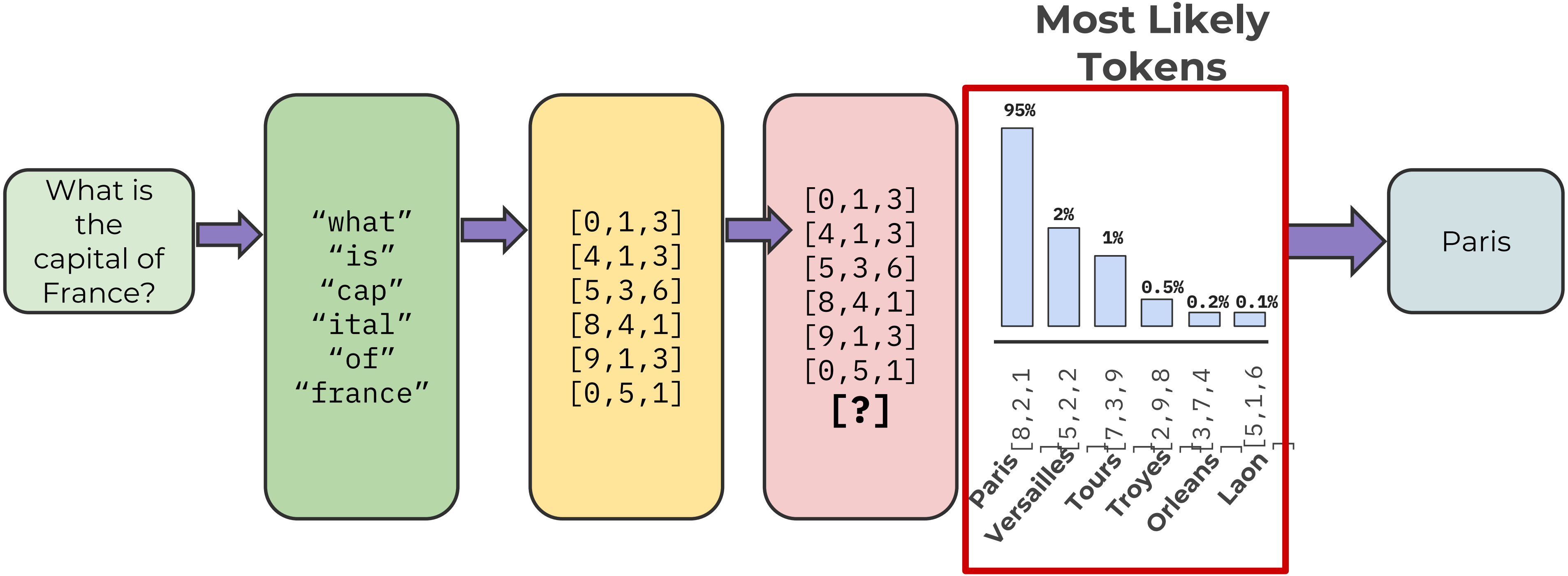
[0,1,3]
[4,1,3]
[5,3,6]
[8,4,1]
[9,1,3]
[0,5,1]

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

Most Likely Tokens

Paris

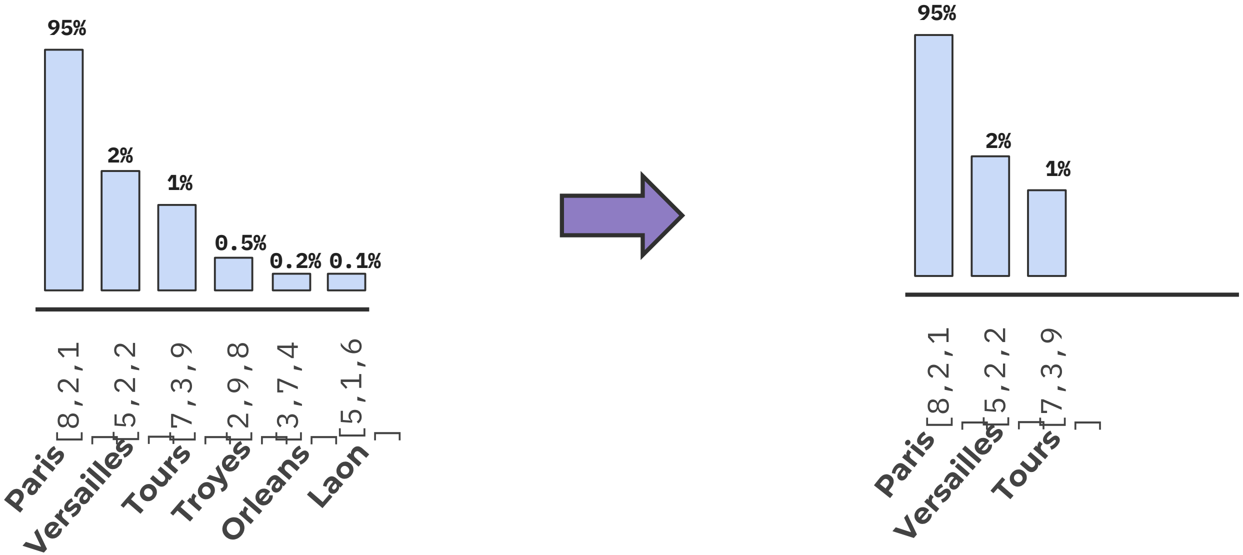
Gemini Python API



- **Top K**
  - This means you would only consider the top K amount of tokens.
  - For example, if K=3, you would only consider the 3 most likely tokens before you sample.

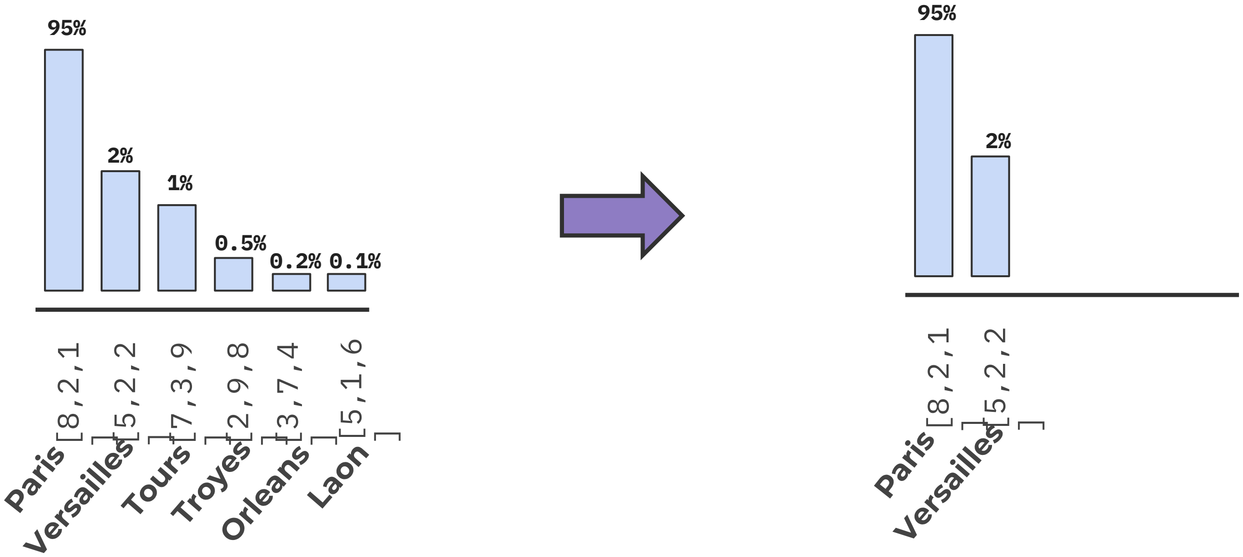
Gemini Python API

- **Top K**
  - With Top K=3



- **Top P**
  - This considers the cumulative probability of the tokens, allowing you to cut-off at a certain cumulative probability.
  - For example, a P = 0.97 would stop considering any tokens once the cumulative probability reaches 97%.

- **Top P**
  - With Top P = 0.97



Gemini Python API

- Let’s explore these configuration parameters with the Python API.