Pierian Training

Embeddings and RAG

Gemini Python API

- **RAG and Embeddings**
  - RAG stands for Retrieval Augmented Generation and makes great use of the model's ability to generate vector embeddings of text.
  - Let's explore this idea with a motivational thought experiment...

Gemini Python API

- **RAG and Embeddings**
  - While Google Gemini is a very powerful model, there will be content it doesn't know about.
  - For example, Google Gemini probably doesn't know about the vacation policies at a specific company.
  - So how could we leverage AI to build a corporate HR assistant?

Gemini Python API

- **RAG and Embeddings**
  - What we need is the ability to “augment” Gemini by providing the model with extra context when asking a question.
  - For example, if we could provide the text about a corporate vacation policy along with our query, then Gemini could read the text and give us a reasonable answer.

Gemini Python API

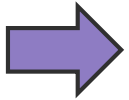
- **RAG and Embeddings**

How many weeks off do employees of ACME Corp get?

Gemini Python API

- **RAG and Embeddings**

How many weeks off do employees of ACME Corp get?



Gemini

Gemini Python API

- **RAG and Embeddings**

How many weeks off do employees of ACME Corp get?

I'm sorry, I can't answer your question about how many weeks off employees of ACME Corp get. I don't have access to internal HR information for specific companies.

However, I can offer some general information about vacation policies in the United States.

Gemini Python API

- **RAG and Embeddings**

Gemini Python API

- **RAG and Embeddings**

How many weeks off do employees of ACME Corp get? Here is some context:

Gemini Python API

- **RAG and Embeddings**

How many weeks off do employees of ACME Corp get? Here is some context:

ACME HR

3 Weeks Vacation per year. With one additional day per year with the company.

Gemini Python API

- **RAG and Embeddings**

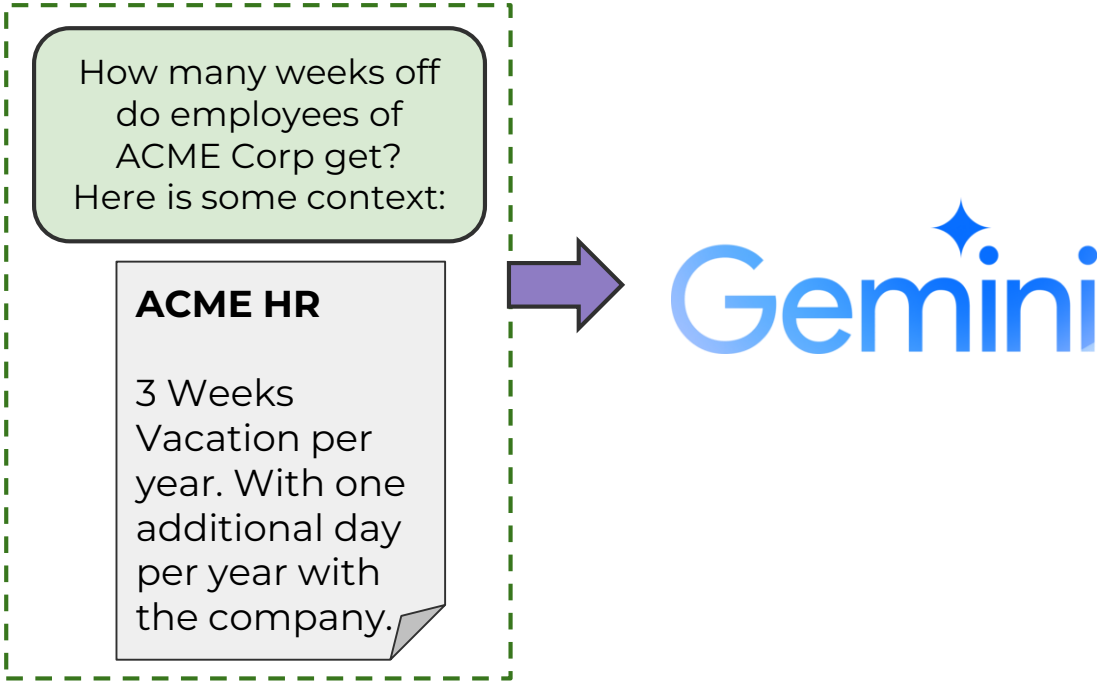
How many weeks off do employees of ACME Corp get? Here is some context:

ACME HR

3 Weeks Vacation per year. With one additional day per year with the company.

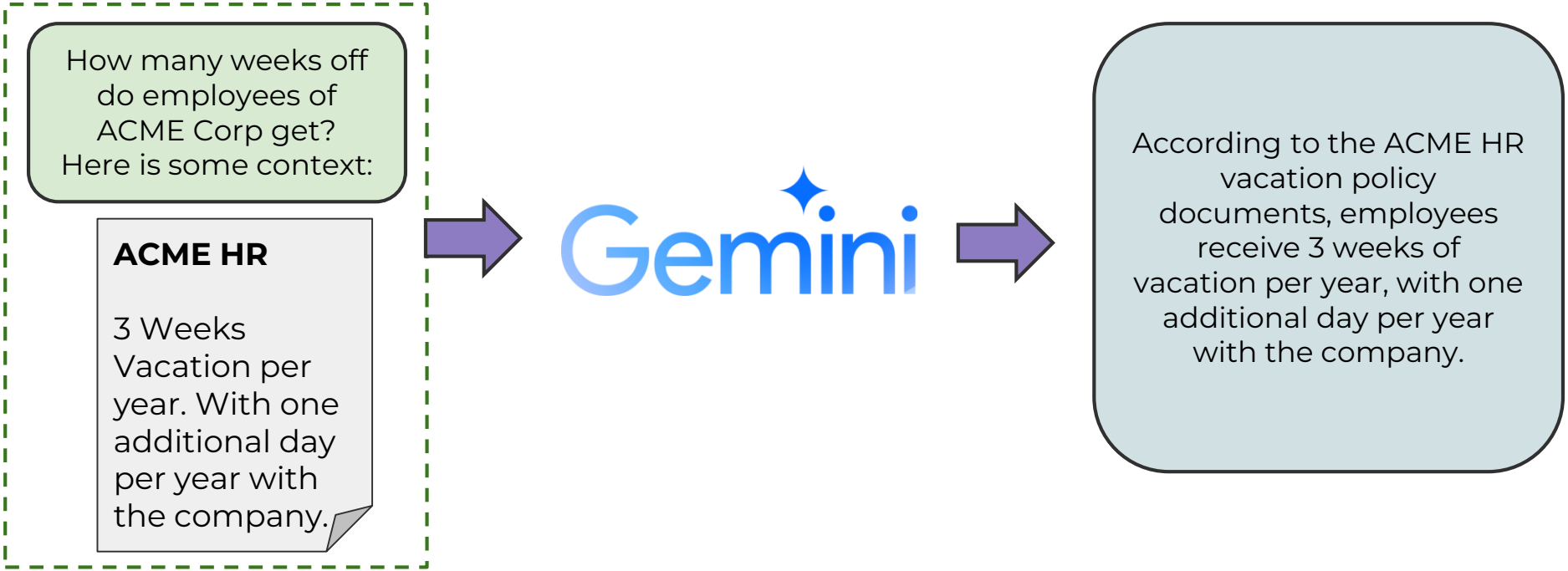
Gemini Python API

- **RAG and Embeddings**



Gemini Python API

- **RAG and Embeddings**

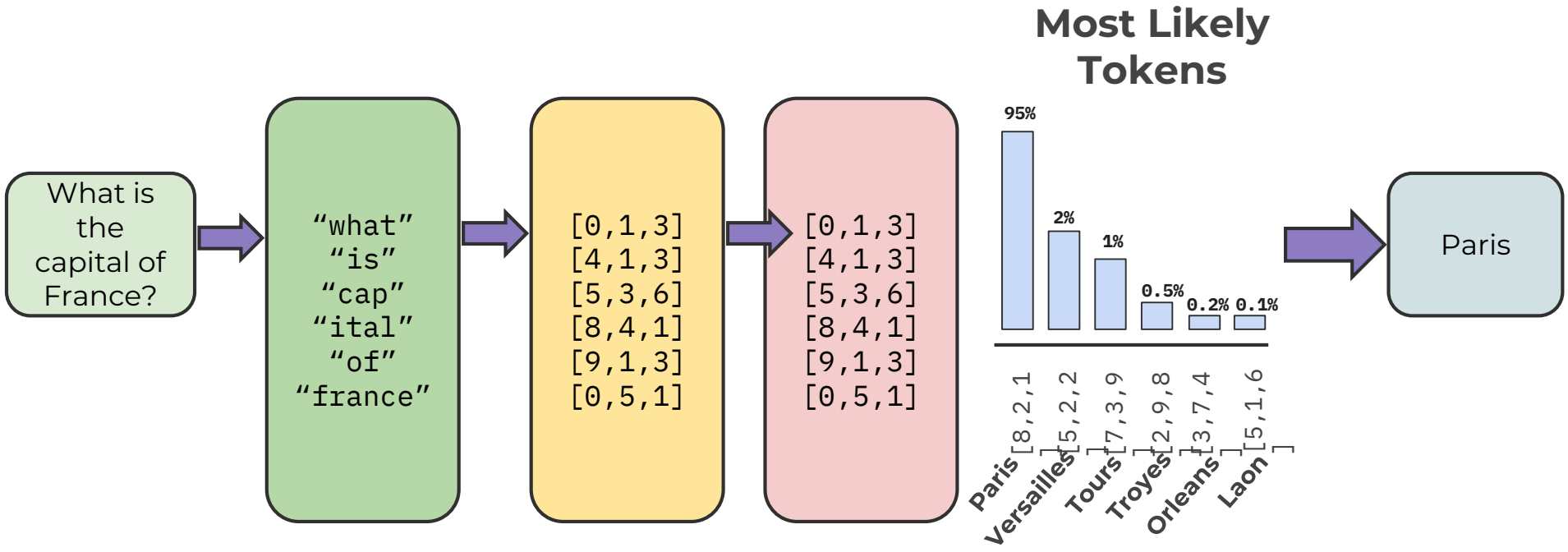


- **RAG and Embeddings**
  - Now we can see that LLMs can answer any question if they have the correct and appropriate context and documents for an answer, this is “augmented generation” because we augmented the original query with the context of the ACME HR policy documents.
  - But how can we **retrieve** this context?

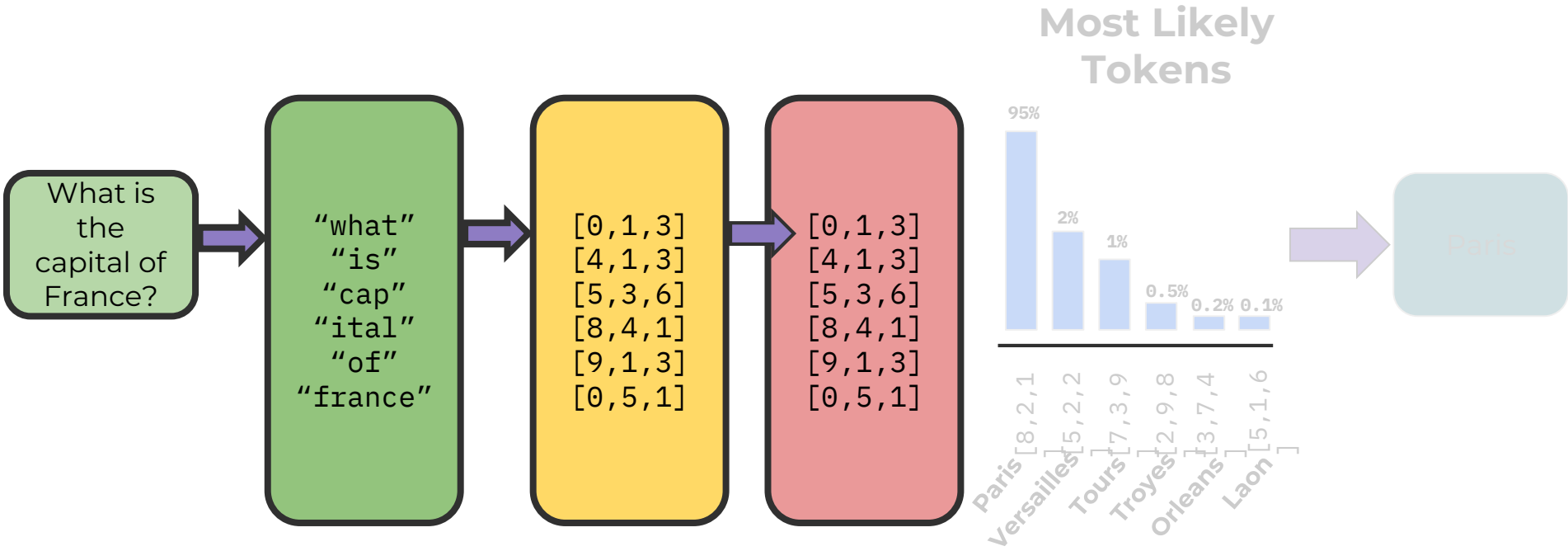
Gemini Python API

- **RAG and Embeddings**
  - Recall our models ability to embed text into vectors.
  - If we had all our text documents stored as vectors and matched to the original text, we could perform vector similarity searches to find the most relevant documents.

Gemini Python API



Gemini Python API

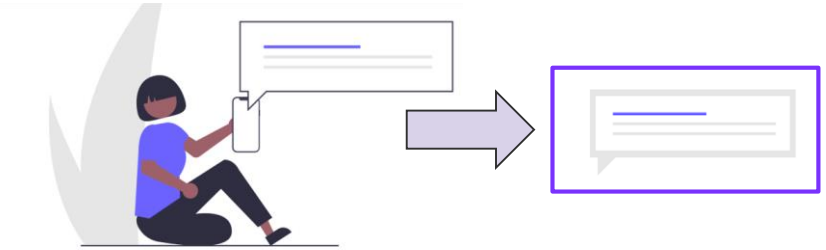


Gemini Python API

- **RAG Process:**

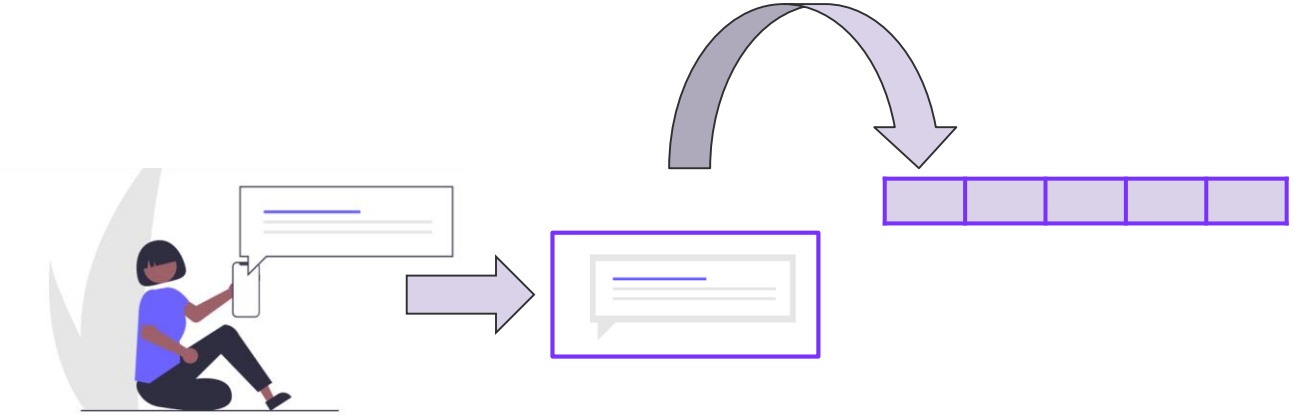
Gemini Python API

- **RAG Process:**

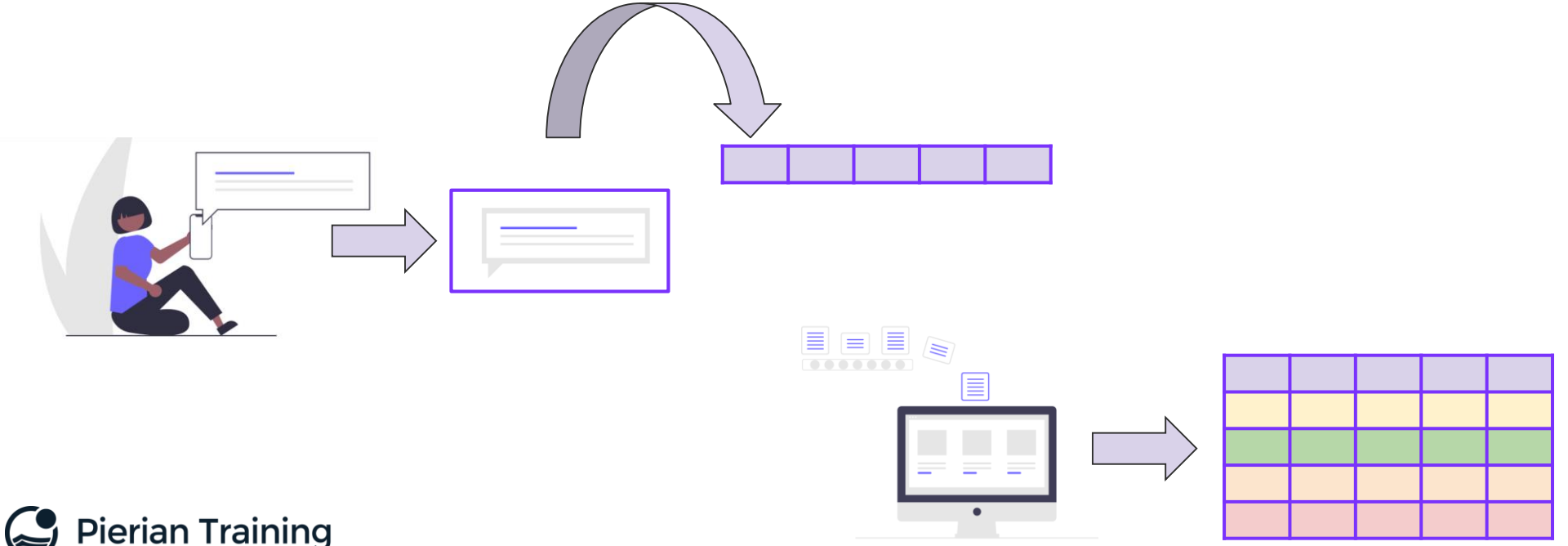


Gemini Python API

- **RAG Process:**

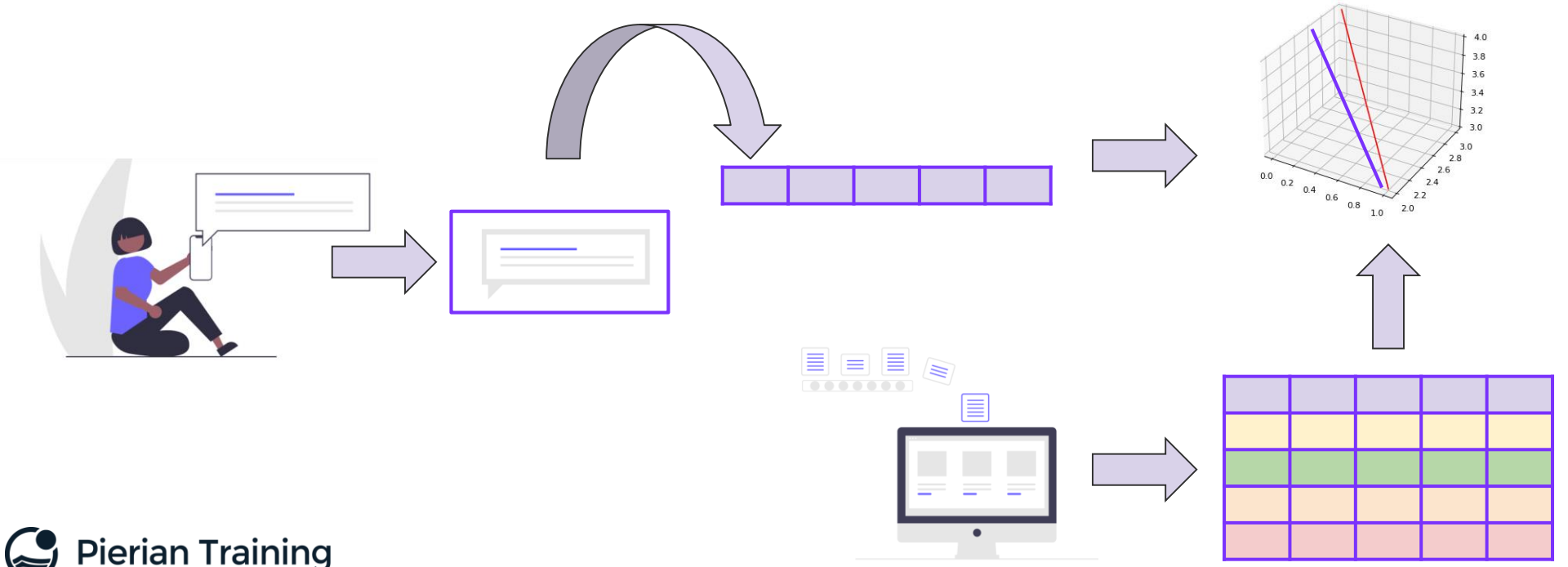


- **RAG Process:**



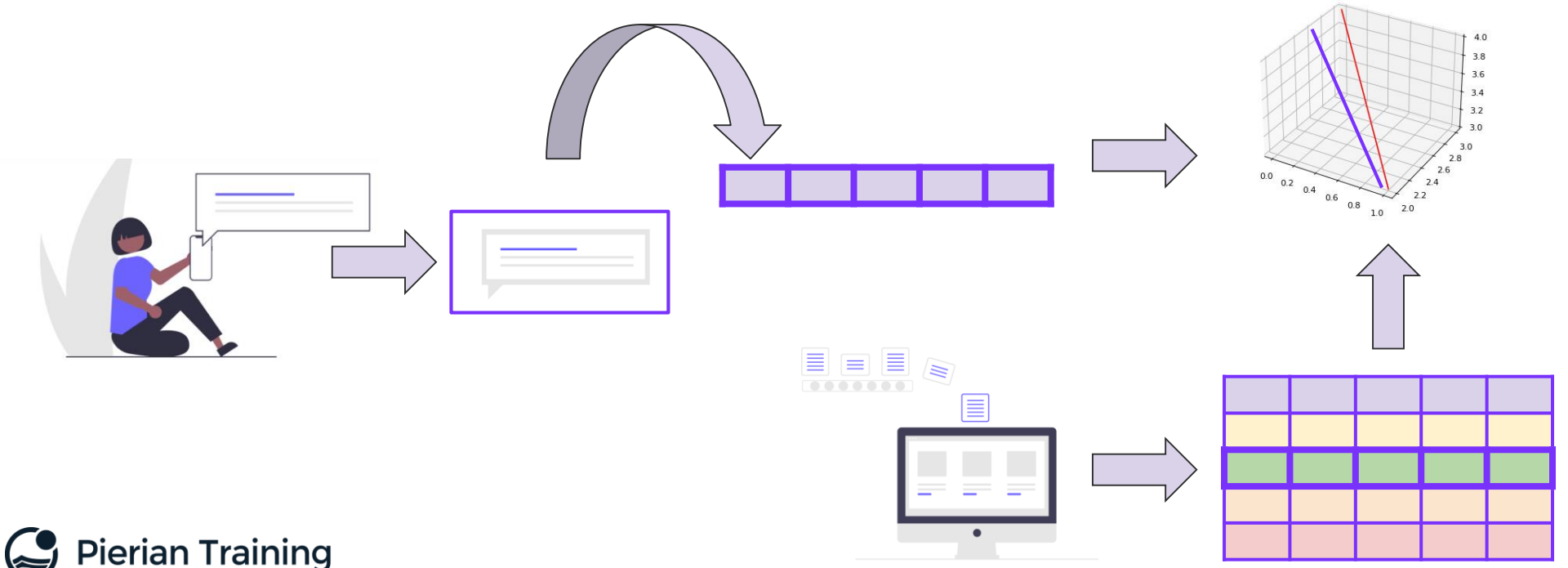
Gemini Python API

- **RAG Process:**



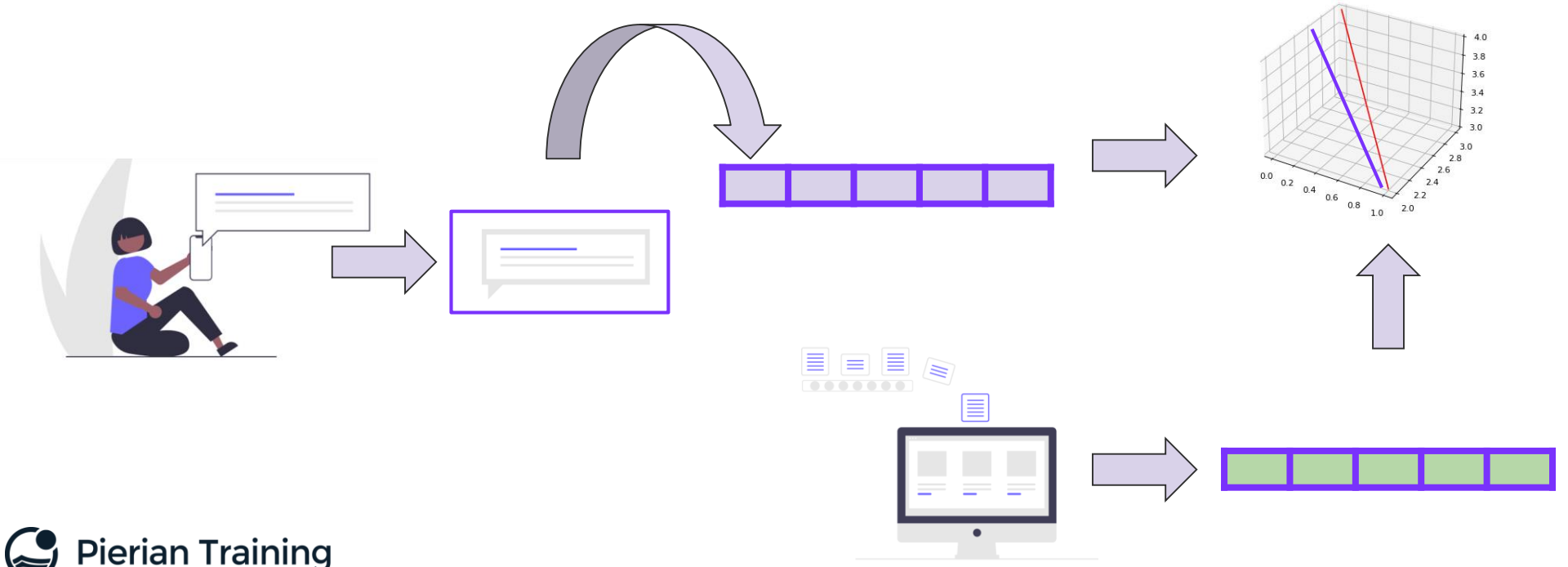
Gemini Python API

- **RAG Process:**



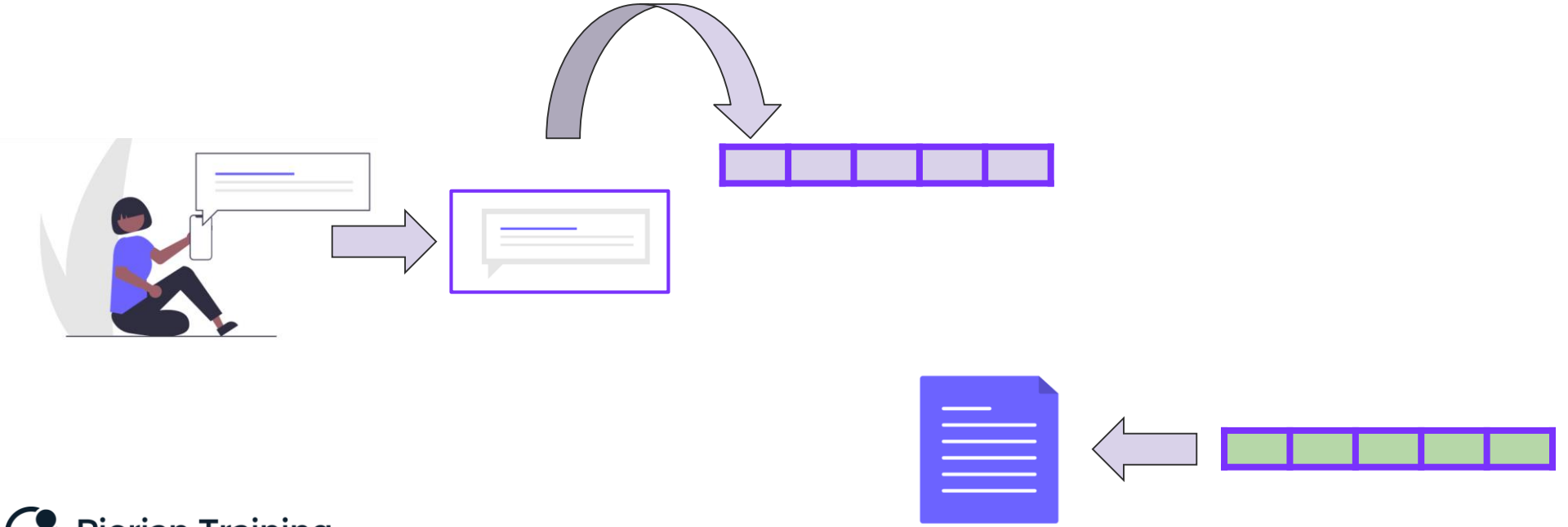
Gemini Python API

- **RAG Process:**



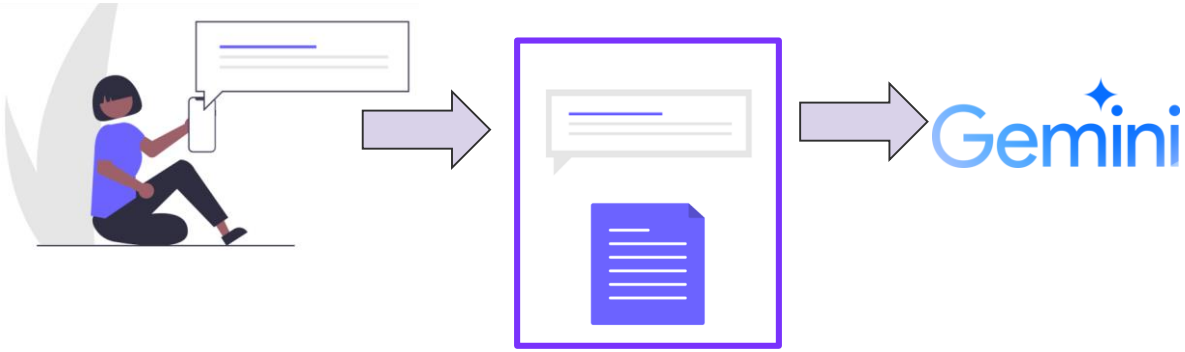
Gemini Python API

- **RAG Process:**



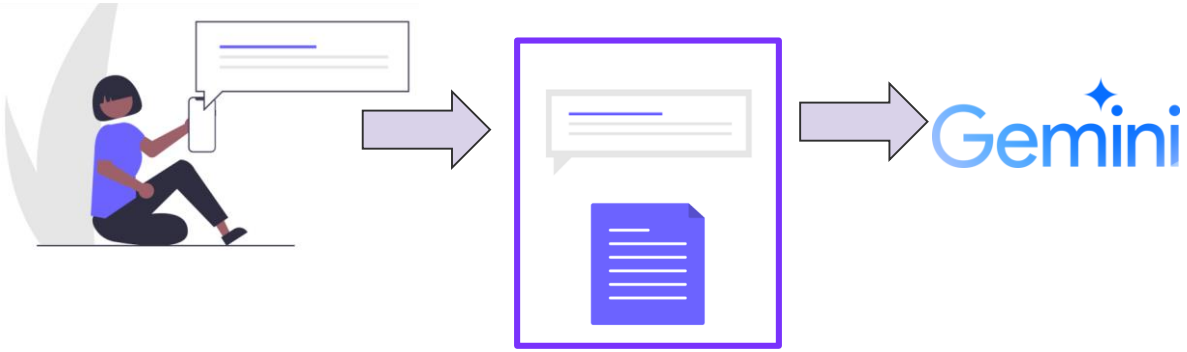
Gemini Python API

- **RAG Process:**



Gemini Python API

- **RAG Process:**



Prompt:

“What can you tell me about vacation policy?”

Context below:

{Insert Document}

Gemini Python API

- **RAG Process Steps**
  - Read in Documents (text, PDF, etc.)
  - Load Embedding Model
  - Embed Documents as Vectors
  - Store Vector Embeddings (Vector Store)
  - Embed new query
  - Perform Similarity Search
  - Augment text generation with original document

- **RAG Process Steps**
  - Remember that as long as you can convert a document into a text string, you should be able to then embed that string and link the connection to the original document.
  - There exists many libraries specific to a document type (e.g. PyPDF2 for PDF files) you should search for relevant libraries or explore unstructured.io for multiple files.

Gemini Python API

- **RAG Process Steps**
  - Let’s explore RAG with Python in the next lecture!

RAG
Part One

RAG
Part Two