LLMs APIS RAG

¿Que es este módulo?

Este módulo integra todo lo aprendido para construir sistemas inteligentes de procesamiento del lenguaje natural usando modelos de lenguaje grandes (LLMs).

Vamos a aprender a:

- Conectarte a APIs de LLMs (OpenAI, Gemini)
- Construir sistemas RAG (Recuperación y Generación Aumentada)
- Trabajar con bases de datos vectoriales
- Integrar búsqueda web y modelos locales
- Crear aplicaciones prácticas de procesamiento de texto

Text Generation

Text Representation

1. Contexto de la tarea

2. Contexto del tono

3. Datos de trasfondo, decumentos e imágenes

4. Descripción detallada de la tarea

5. Ejemplos

6. Historial de conversación

7. Descripción inmediata de la tarea o solicitud

8. Pensar paso a paso / respirar profundo

9. Formato de salida

10. Respuesta prellenada (si aplica)

Usuario

Estarás actuando como un coach de carrera de IA llamado Joe creado por la empresa AdAstra Careers. Tu objetivo es dar consejos de carrera a los usuarios. Estarás respondiendo a usuarios que están en el sitio web de AdAstra y que se confundirán si no respondes en el personaje de Joe.

Debes mantener un tono amigable de servicio al cliente.

Aquí está el documento de orientación profesional que debes referenciar al responder al usuario: <guide>[[DOCUMENT]]</guide>

Aquí están algunas reglas importantes para la interacción:

- Mantente siempre en personaje, como Joe, una IA de AdAstra careers
- Si no estás seguro de cómo responder, di "Lo siento, no entendí eso. ¿Podrías repetir la pregunta?"
- Si alguien pregunta algo irrelevante, di "Lo siento, soy Joe y doy consejos de carrera. ¿Tienes alguna pregunta sobre carrera con la que pueda ayudarte hoy?"

Aquí hay un ejemplo de cómo responder en una interacción estándar:
<example> Usuario: Hola, ¿cómo fuiste creado y qué haces? Joe: ¡Hola! Mi nombre es Joe, y fui creado por AdAstra Careers para dar consejos de carrera. ¿En qué puedo ayudarte hoy? </ example>

Aquí está el historial de conversación (entre el usuario y tú) previo a la pregunta. Podría estar vacío si no hay historial: <history> [[HISTORY]] </history> Aquí está la pregunta del usuario: <question> [[QUESTION]] </question>

¿Cómo respondes a la pregunta del usuario?

Piensa en tu respuesta primero antes de responder.

Pon tu respuesta en etiquetas <response></response>.

Asistente (prerellenado)

<response>

```python
from google import genai

client = genai.Client()

response = client.models.generate_content(
    model="gemini-2.5-flash",
    contents="Explain how AI works in a few words",
)

print(response.text)
```

```python
import anthropic

anthropic.Anthropic().messages.create(
    model="claude-sonnet-4-20250514",
    max_tokens=1024,
    messages=[
        {"role": "user", "content": "Hello, world"}
    ]
```

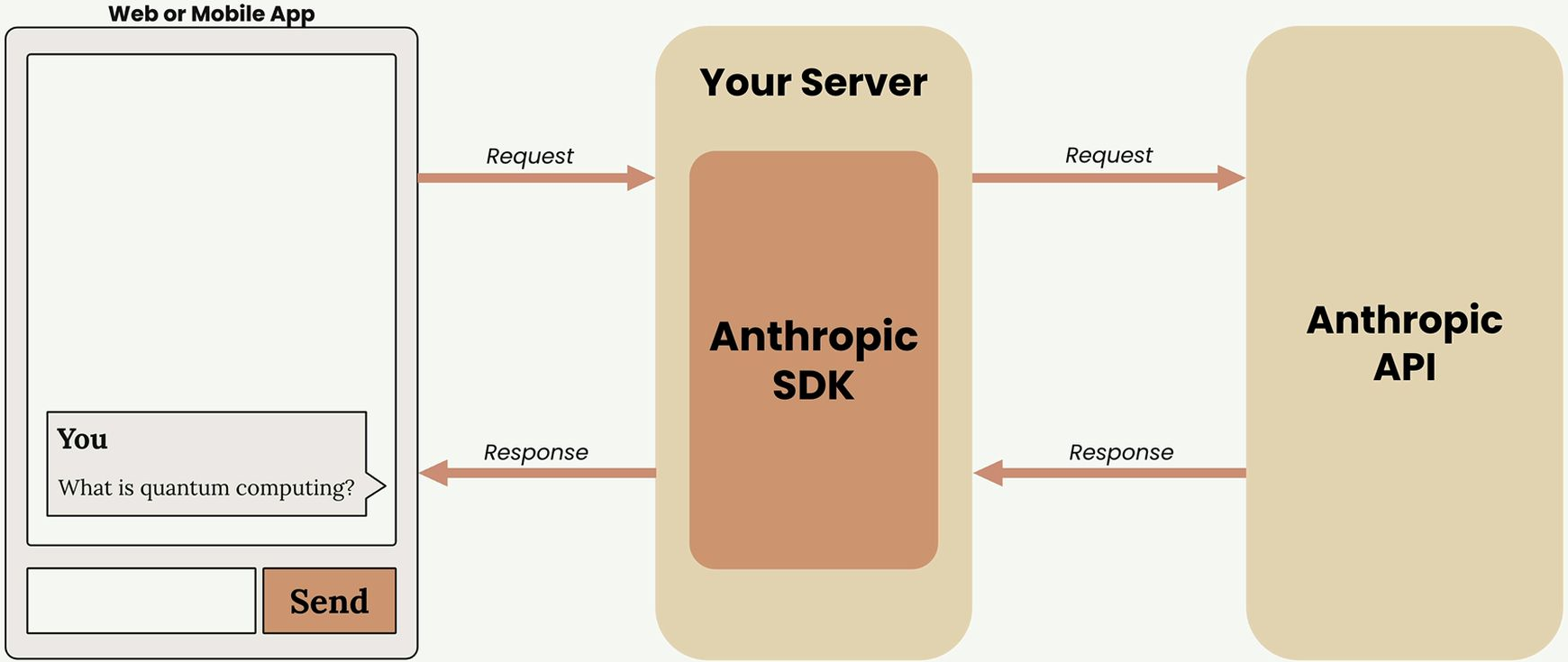
```python
from openai import OpenAI
client = OpenAI()

response = client.responses.create(
    model="gpt-5",
    input="Write a short bedtime story about a unicorn."
)

print(response.output_text)
```

¿Qué es una API?

Una ****API**** (Application Programming Interface) es una interfaz que permite que dos sistemas se comuniquen entre sí. En nuestro caso, nos permite enviar texto a un modelo de lenguaje (como GPT-4) y recibir una respuesta procesada.


Web or Mobile App
You
What is quantum computing?
Send
Request
Response
Your Server
Anthropic SDK
Request
Response
Anthropic API

Parámetros Configurables en LLMs

Temperature (0-2): Controla la aleatoriedad de las respuestas
- `0.0` = Deterministico, siempre responde igual (ideal para extraer datos)
- `0.7` = Balance entre creatividad y consistencia
- `1.5-2.0` = Muy creativo y variado

Model: Que modelo usar (calidad vs costo vs velocidad)
- `gpt-4o-mini` = Rapido, economico
- `gpt-4o` = Mas potente, mejor razonamiento

Google Search API

Scrape Google and other search engines from our fast, easy, and complete API.

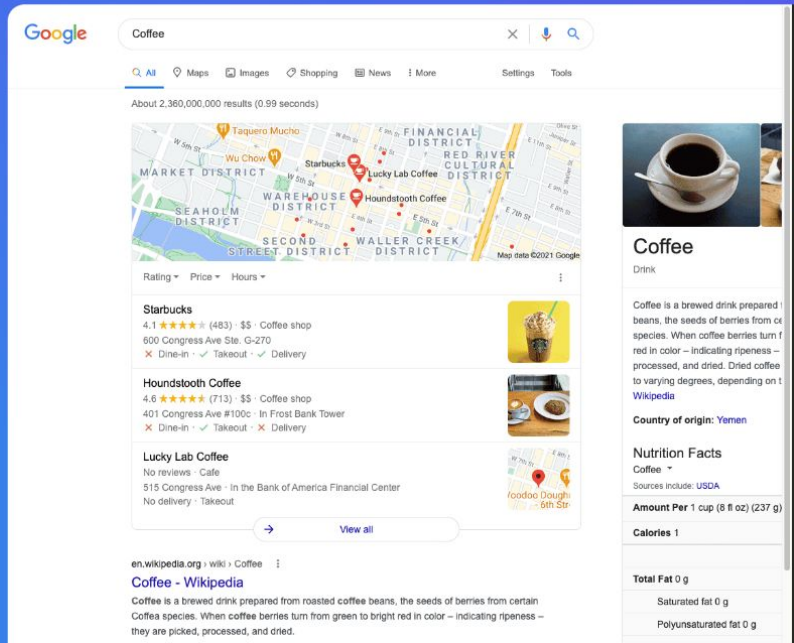
Search Query

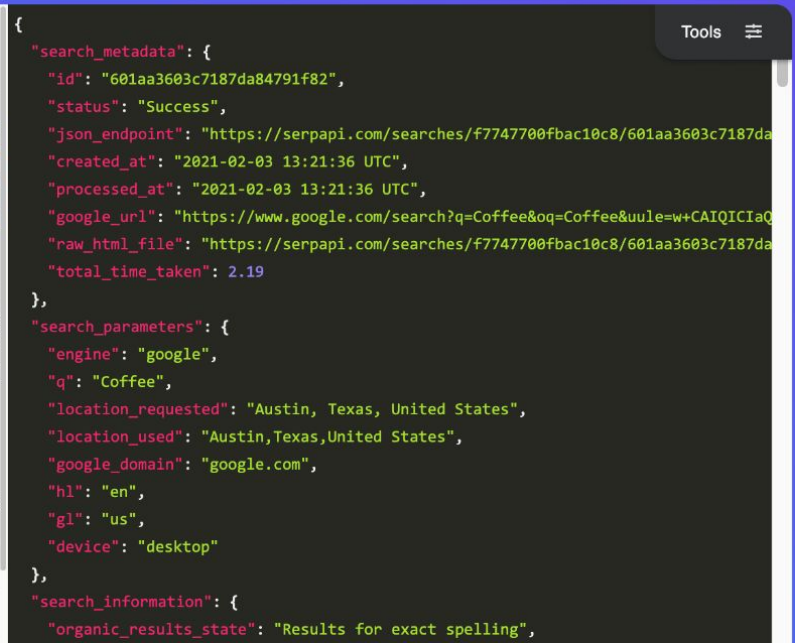
Coffee

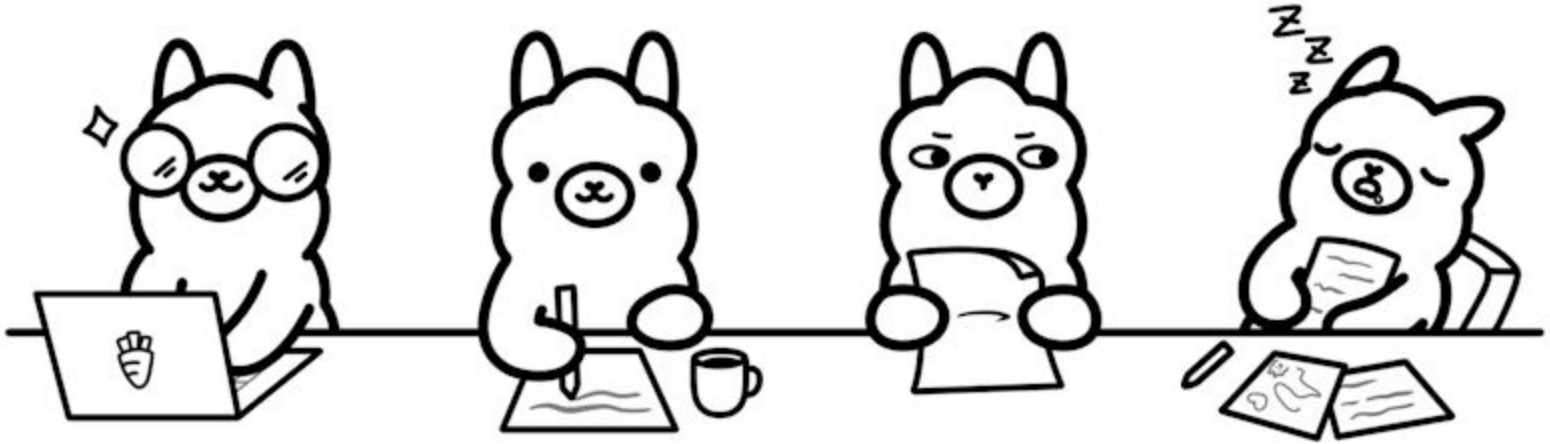
Location

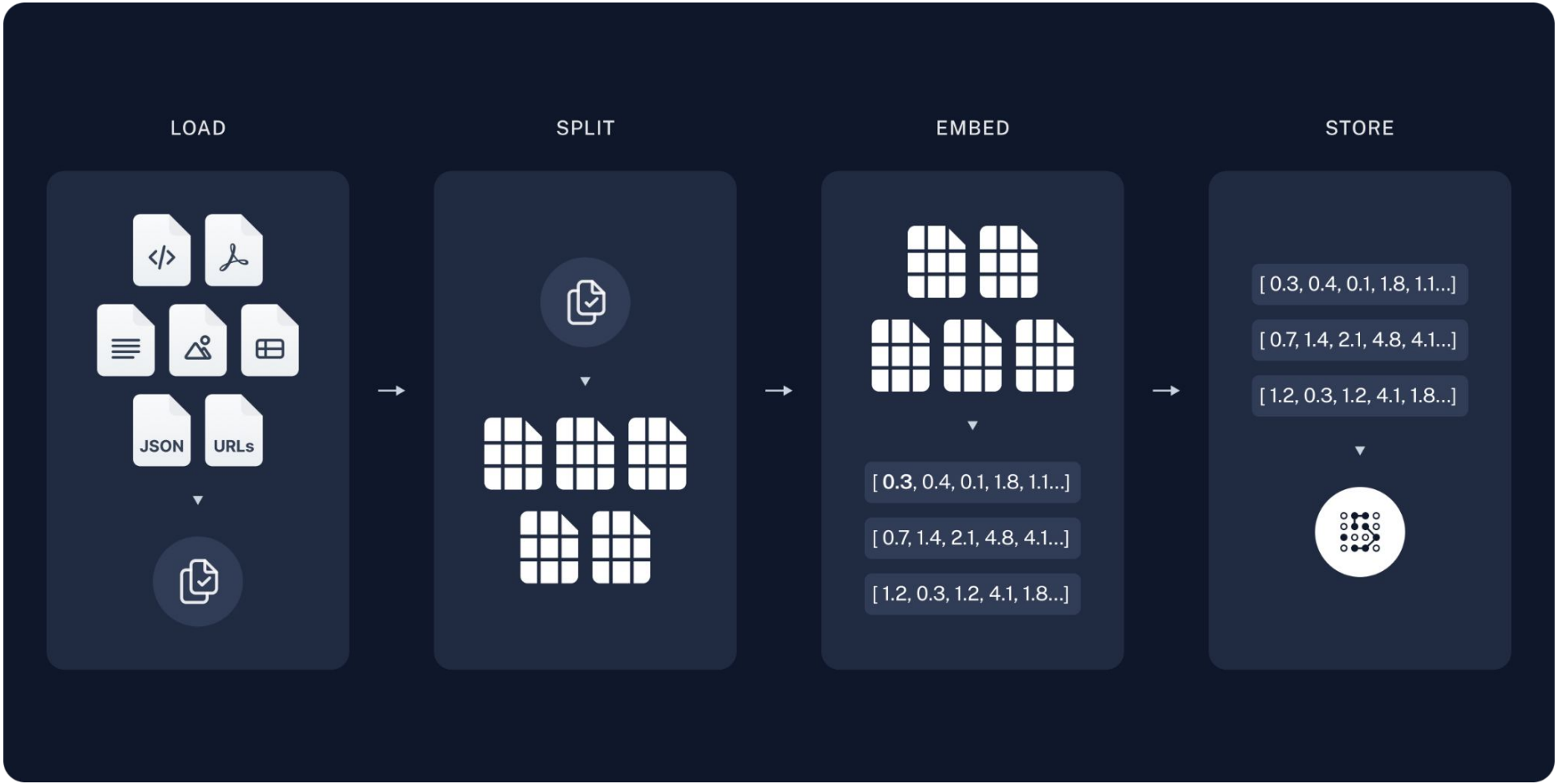
Austin, Texas, United States

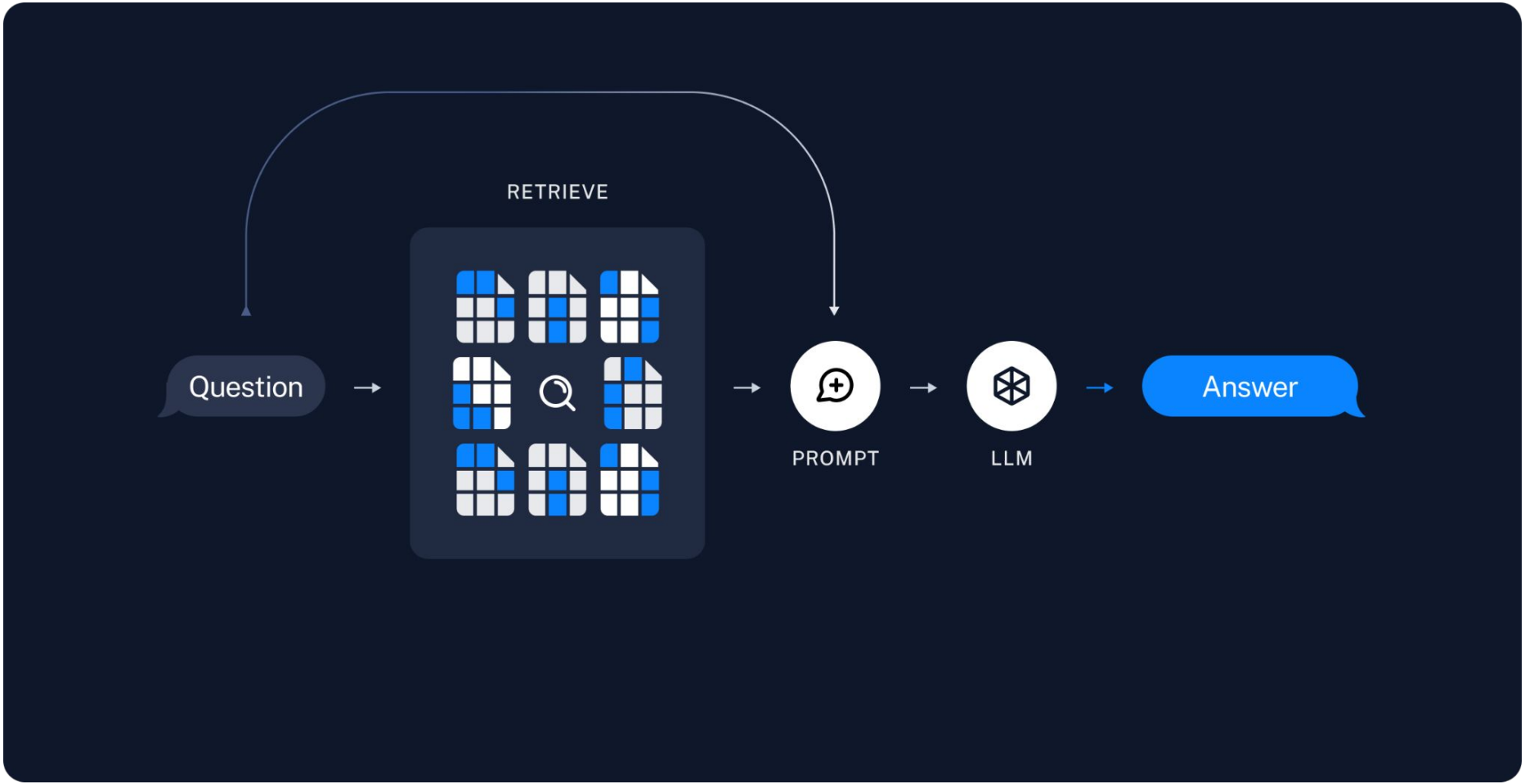
TEST SEARCH

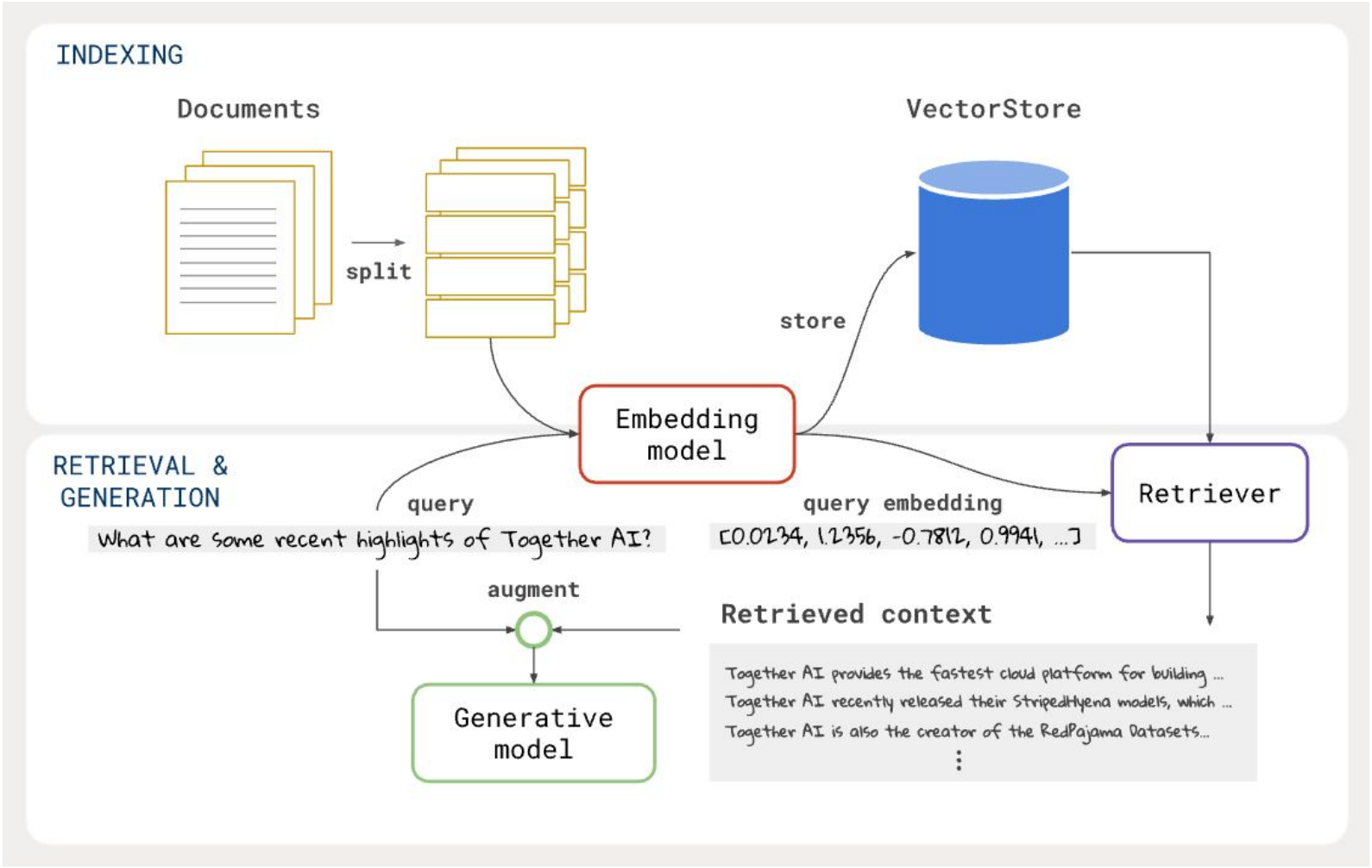






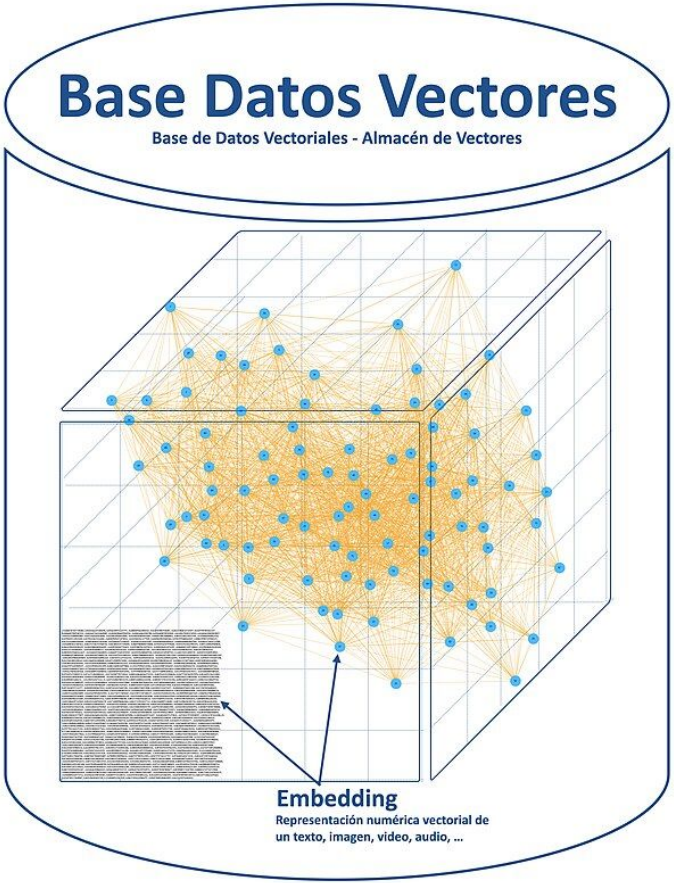

LOAD
SPLIT
EMBED
STORE
JSON
URLs
[0.3, 0.4, 0.1, 1.8, 1.1...]
[0.7, 1.4, 2.1, 4.8, 4.1...]
[1.2, 0.3, 1.2, 4.1, 1.8...]
[0.3, 0.4, 0.1, 1.8, 1.1...]
[0.7, 1.4, 2.1, 4.8, 4.1...]
[1.2, 0.3, 1.2, 4.1, 1.8...]


RETRIEVE
Question
PROMPT
LLM
Answer


INDEXING
Documents
split
VectorStore
store
Embedding model
RETRIEVAL & GENERATION
query
What are some recent highlights of Together AI?
query embedding
[0.0234, 1.2356, -0.7812, 0.9941, ...]
Retriever
augment
Retrieved context
Together AI provides the fastest cloud platform for building ...
Together AI recently released their StripedHyena models, which ...
Together AI is also the creator of the RedPajama Datasets...
Generative model

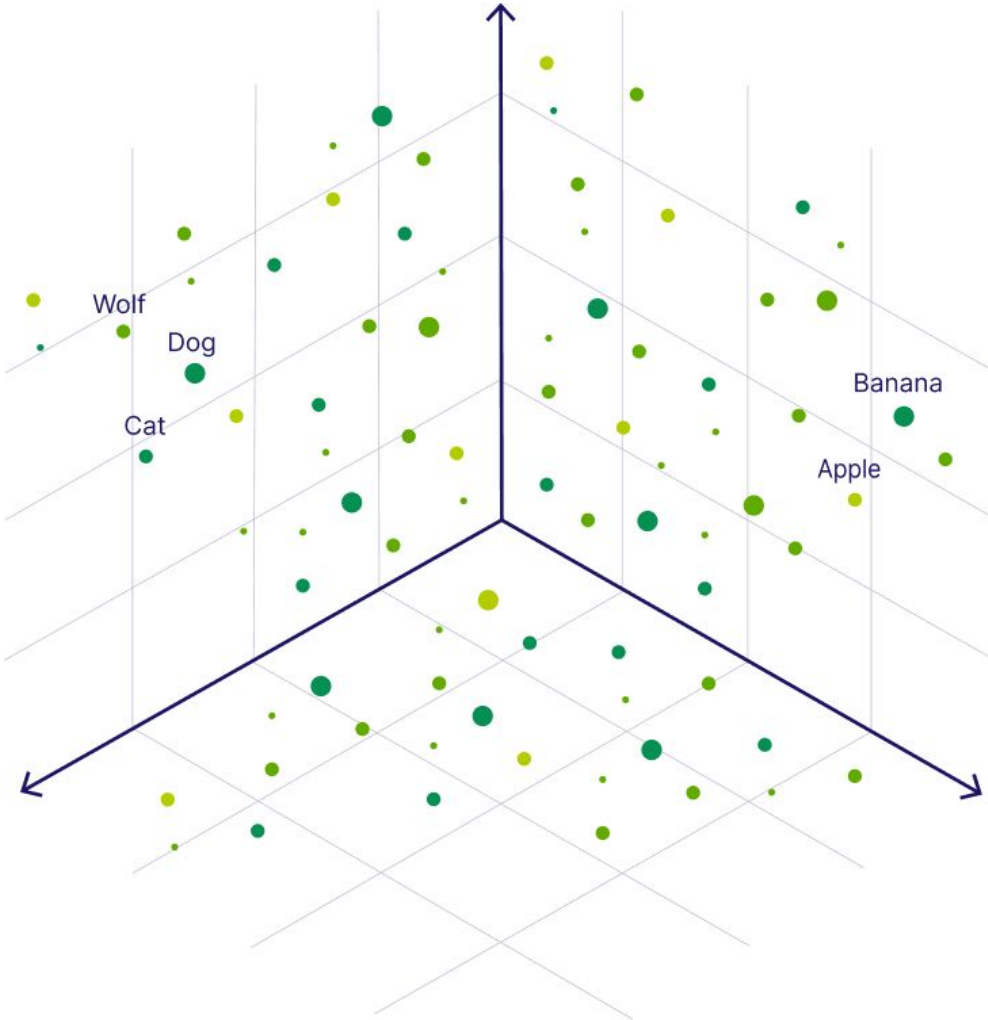
Base Datos Vectores

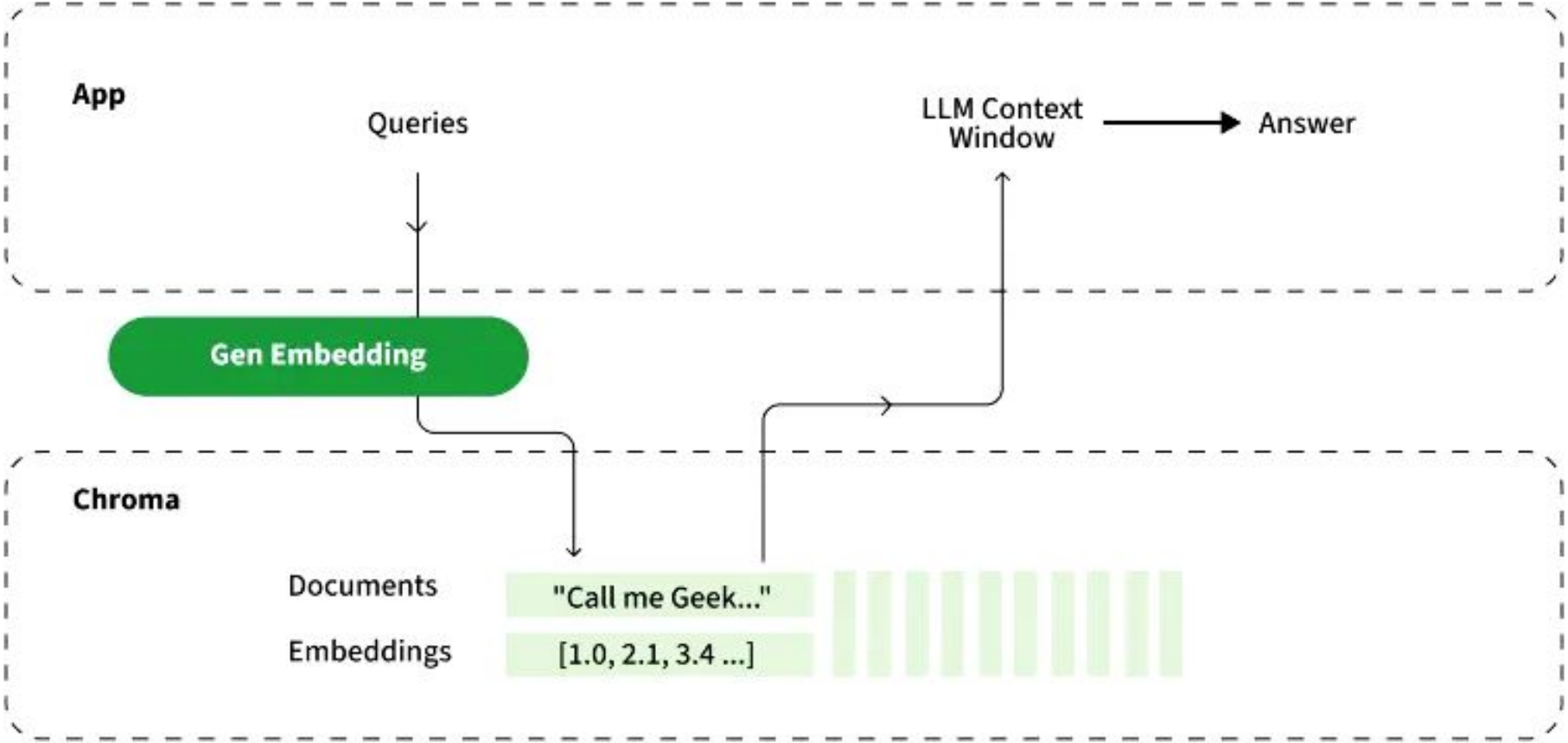
Base de Datos Vectoriales - Almacén de Vectores



Embedding

Representación numérica vectorial de un texto, imagen, video, audio, ...


Wolf
Dog
Cat
Banana
Apple


App
Queries
LLM Context Window
Answer
Gen Embedding
Chroma
Documents
"Call me Geek..."
Embeddings
[1.0, 2.1, 3.4 ...]