RETOS AL CONSTRUIR UNA APLICACIÓN OPEN SOURCE DE IA

Pythonistas

SPEAKERS

@JACKBRAVO (X)
LINKEDIN.COM/IN/JACKBRAVO

JORGE ANTOLIN
BIT.LY/JORGE-MEXICO-LINKEDIN

FABIAN VELAZQUEZ
BIT.LY/FABIAN-LINKEDIN

UN SISTEMA RAG

¿QUÉ ES UN SISTEMA RAG?

¿CÓMO ENSEÑAR COSAS NUEVAS A LA IA?

Pythonistas

DOS METODOS

- Copy & Paste
- Fine tuning

Fine Tuning

Agregar más conocimiento a los pesos del modelo.

Requiere:

- Buenos datos, duh, **muchos datos**
- Hardware con GPU+CPU, mucho RAM, mucho SSD. O rentar. Al entrenar
- Un par de días / semanas / meses? cada vez que lo hagas
- Hardware con GPU+CPU, mucho RAM. O rentra. Al correr.

Copy & Paste

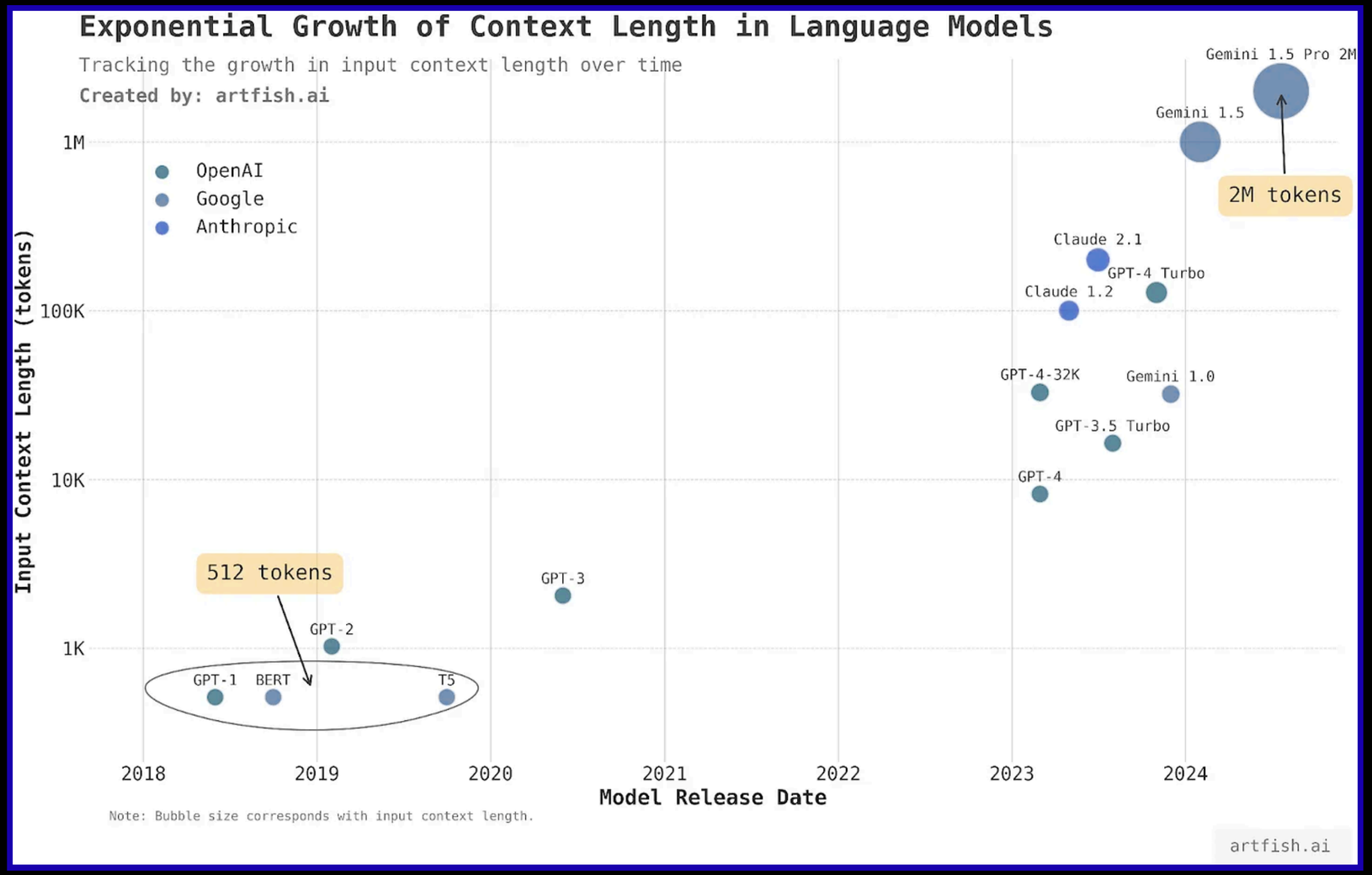
Pythonistas

```
# Employee handbook

The in-line promotion process is designed to
acknowledge employees who have gained additional
skills, abilities, and competencies over a defined
period and are ready to step into a position with
greater autonomy and organizational contribution.

An in-line promotion may occur in the following
circumstances:

- When an employee, in their current role, gains
additional knowledge, skills, competency, or
responsibilities through an established family of
positions with increased scope, complexity, and/or
impact to the business.

- When an employee demonstrates consistent and
sustainable performance beyond their current
position.
```


Input prompt

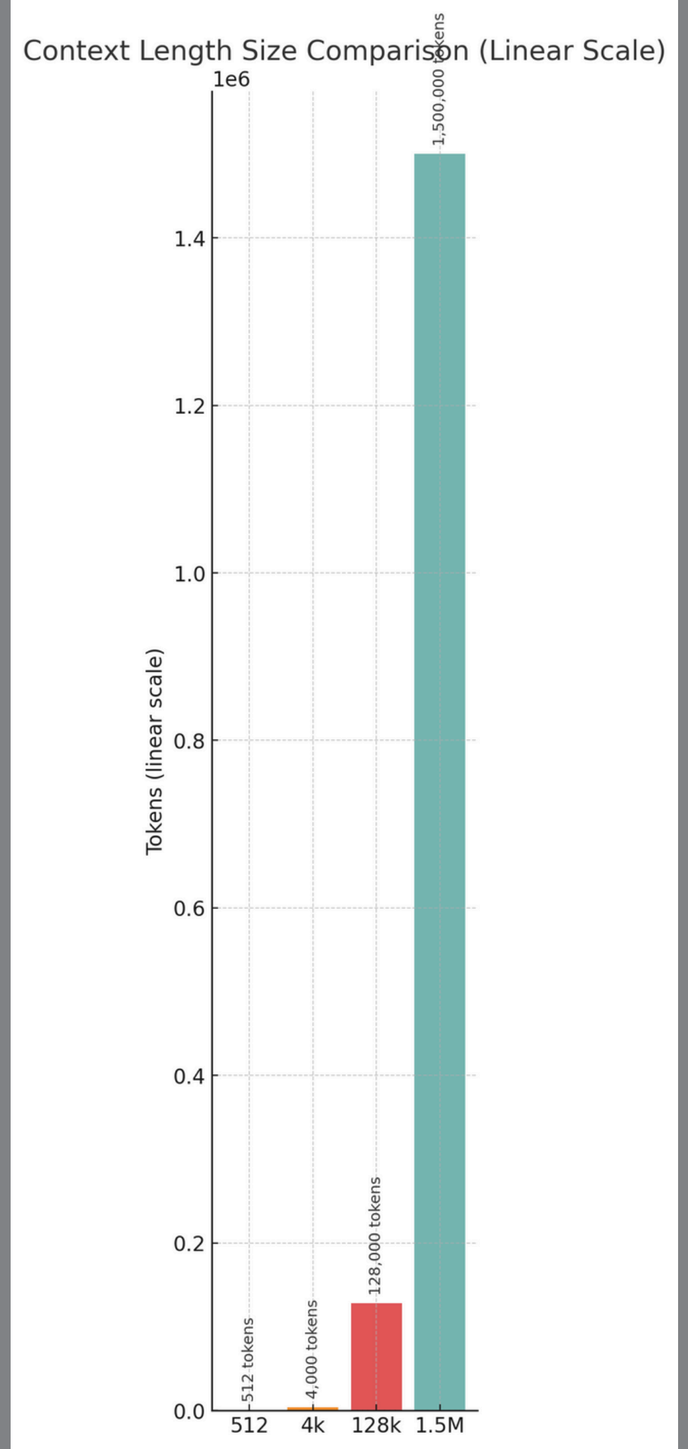
```
Context:

## Employee handbook
...

Given the context, answer the following question:

When can I apply for a promotion?
```

¿QUE TANTO PUEDES PEGAR?



Tamaños de tokens

- 512 tokens: 1 página
- 1k tokens: 2.5 páginas
- 4k tokens: Como 10 páginas
- 32k tokens: Como 100 páginas.
- 128k tokens: Como 400 páginas.
- 1M tokens: 2,500 páginas 😱



PERO... ¿Y QUÉ TAL SI TENEMOS **MÁS** DATOS?
... O QUEREMOS CITAS Y REFERENCIAS **CONFIABLES**?

Necesitas un **RAG?**

Un sistema **RAG** (Retrieval-Augmented Generation), es un enfoque que le da "superpoderes" a un modelo de lenguaje (LLM) permitiéndole consultar información específica antes de responder.

Una base de datos

Con búsquedas

¿COMO FUNCIONA?

RECUPERACIÓN (RETRIEVAL)
BÚSQUEDA

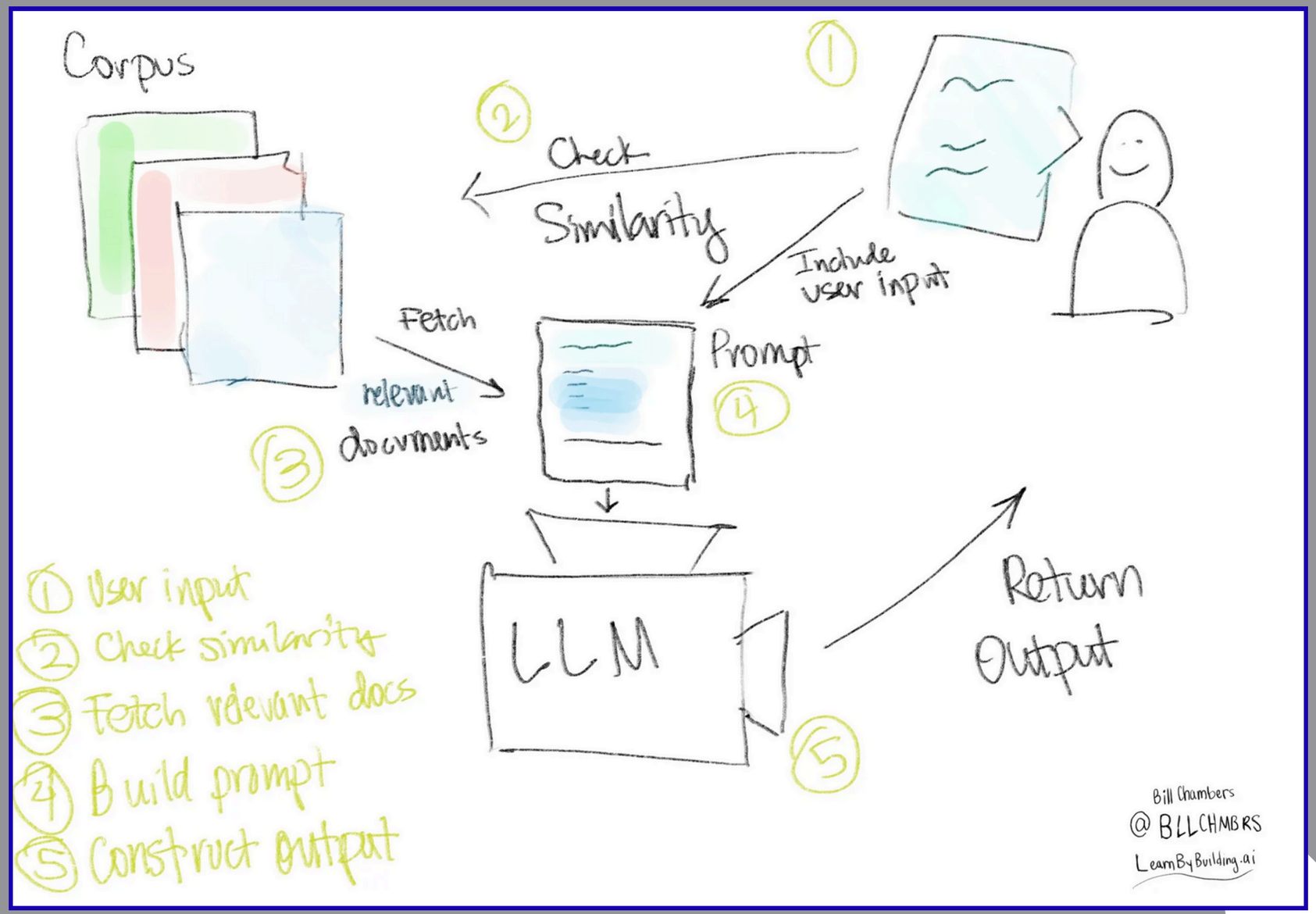
AUMENTADA (AUGMENTED)
-MEJORADA-

GENERACIÓN (GENERATION)
-DE RESPUESTAS-

GENERACIÓN DE RESPUESTAS MEJORADA POR BÚSQUEDAS

RAG Prompt

```
Context:
{context}

Instructions:
You are an assistant for question-answering tasks. Use the
following pieces of retrieved context to answer the question. If
you don't know the answer, just say that you don't know.

Question:
{question}
```


Corpus
1
2
Check
Similarity
Include
user input
Fetch
relevant
documents
3
Prompt
4
LLM
5
Return
Output
1 User input
2 Check similarity
3 Fetch relevant docs
4 Build prompt
5 Construct output
Bill Chambers
@ BLLCHMBRS
LearnByBuilding.ai

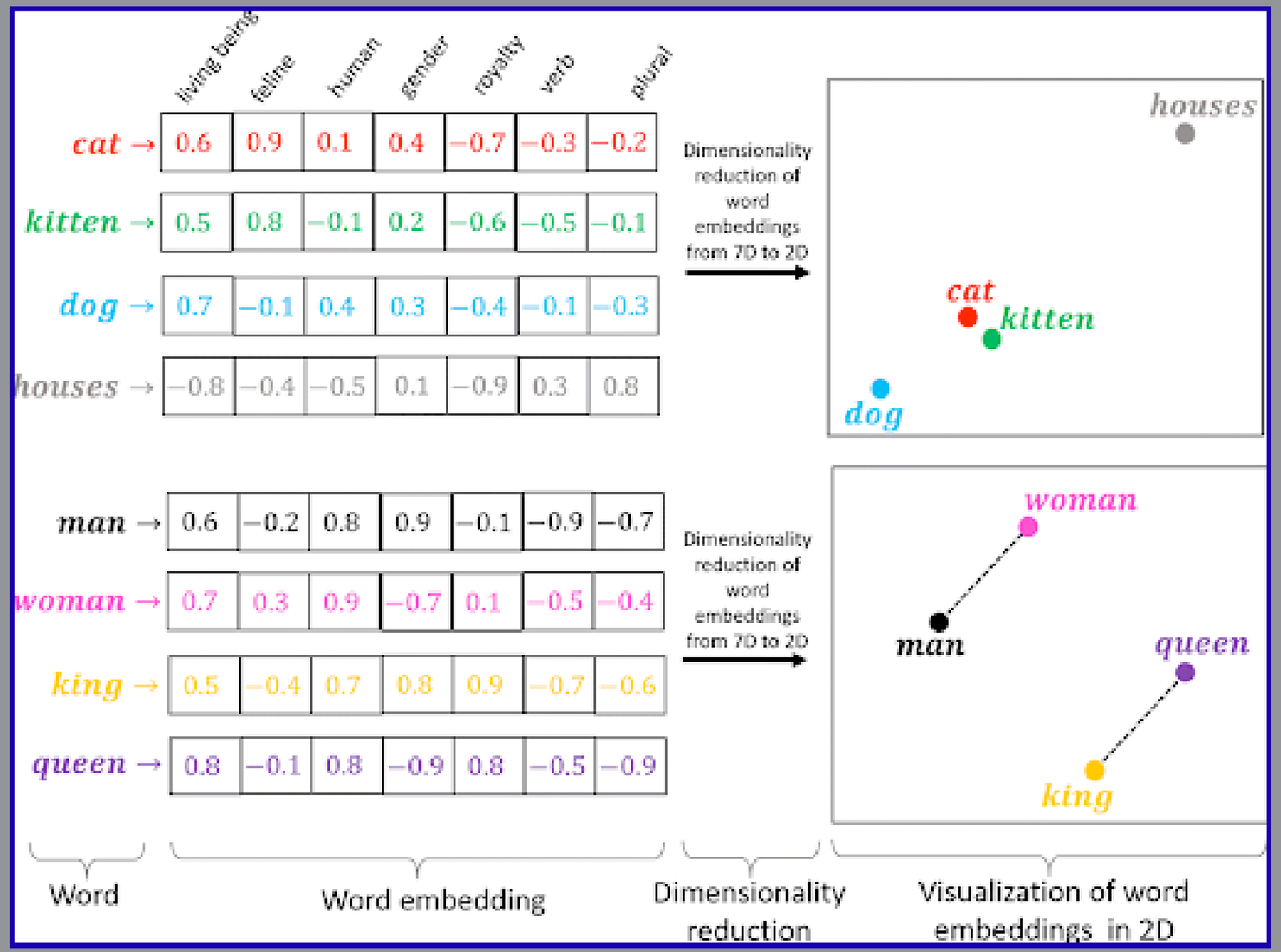
¿CÓMO BUSCAR LOS MEJORES DOCS?

- **BÚSQUEDA POR KEYWORDS (ALA GOOGLE):**
  - FULL TEXT SEARCH (SOLR, ELASTICSEARCH, POSTGRES)
  - FASTER AND SMARTER THAN LIKE ‘%?%’
  - TF-IDF (QUE TAN POPULAR ES UN TÉRMINO EN EL DOCUMENTO, PERO RARO EN EL UNIVERSO DE DOCUMENTOS)

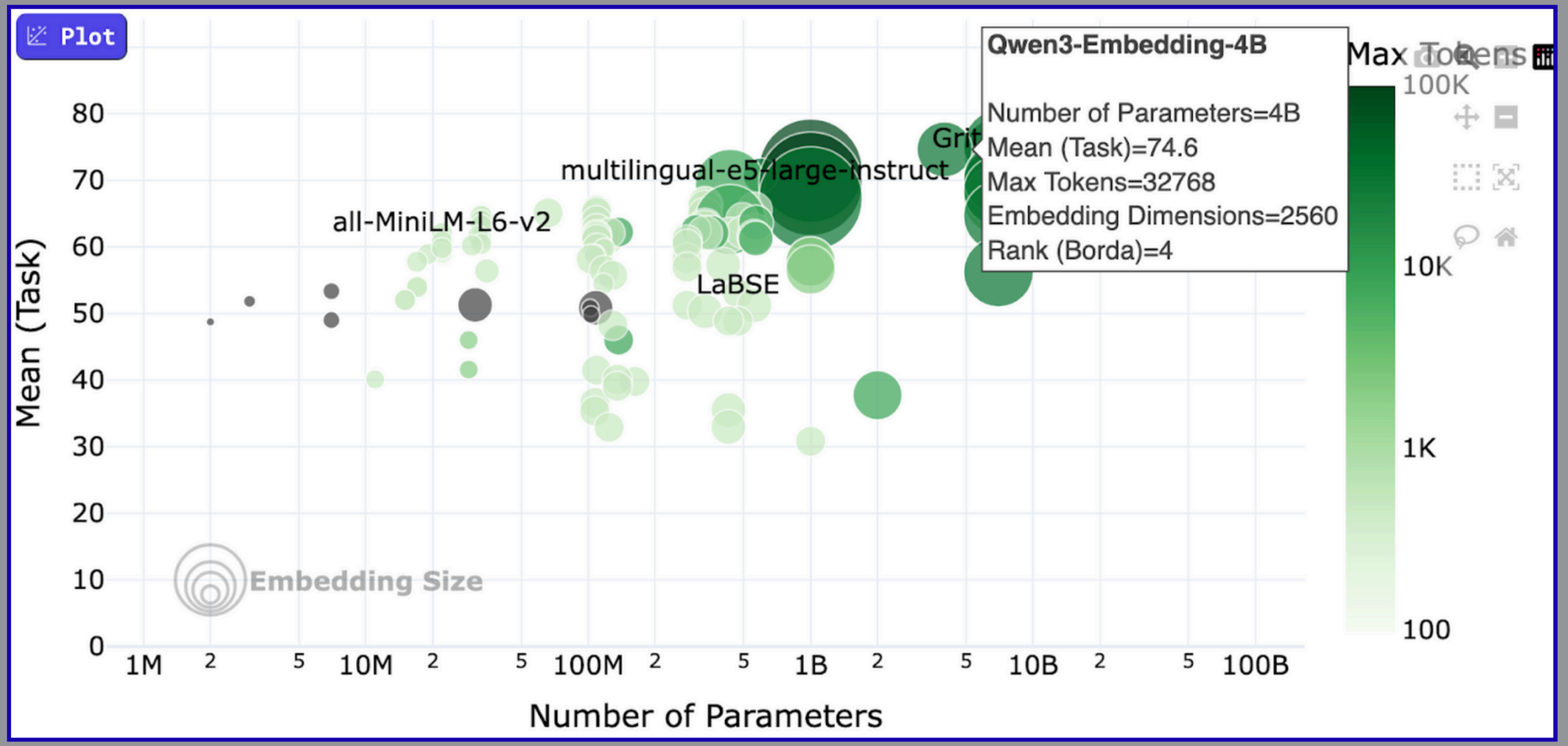
- **BÚSQUEDA POR SIMILARIDAD:**
  - USA UN MODELO DE IA DE EMBEDDINGS
  - BÚSCA DE ACUERDO AL SENTIDO DEL TEXTO, NO SUS KEYWORDS
  - ORDER BY ARRAY_DISTANCE(VEC_COLUMN, USER_VEC)

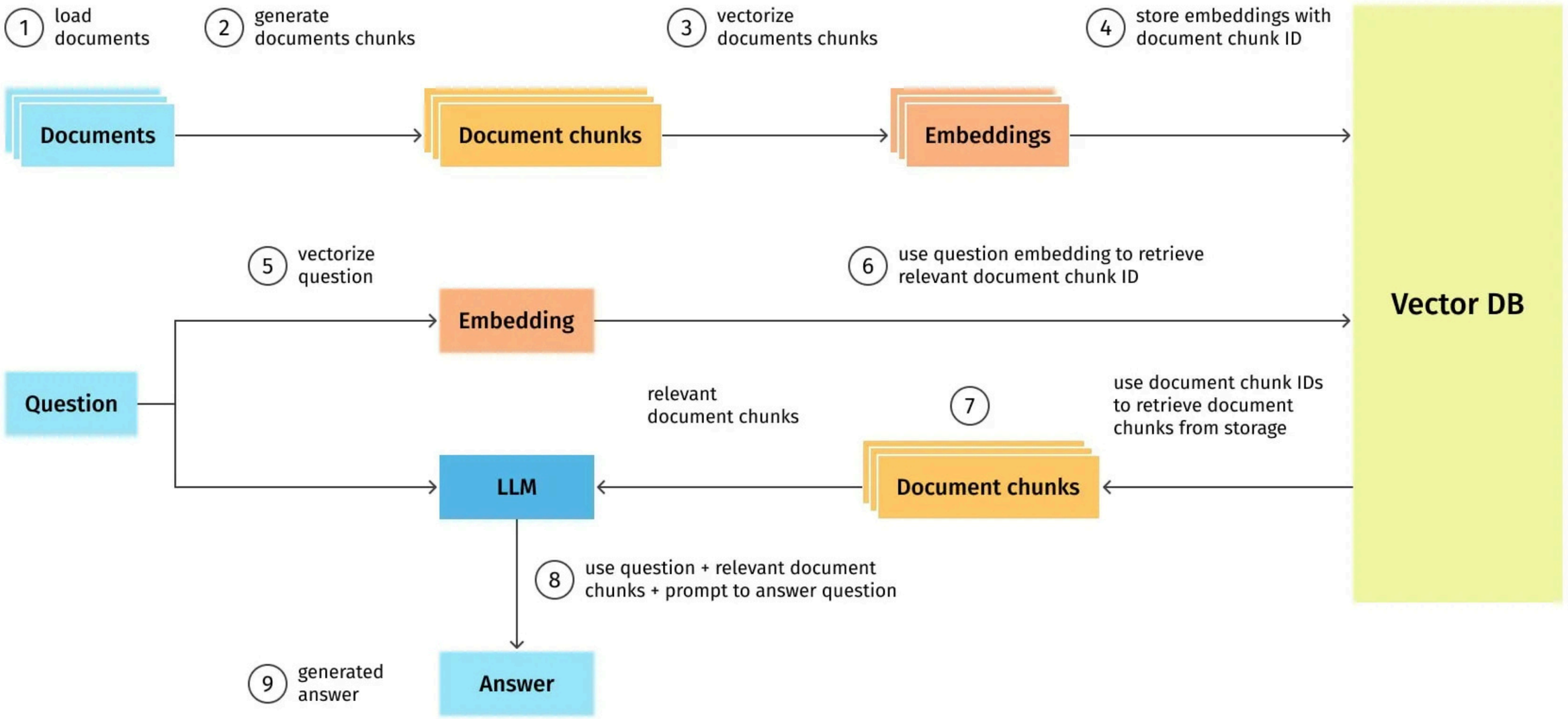
APRENDE MÁS SOBRE EMBEDDINGS EN ESTA GUÍA VISUAL EN INGLÉS

¿QUE ES UN EMBEDDING?

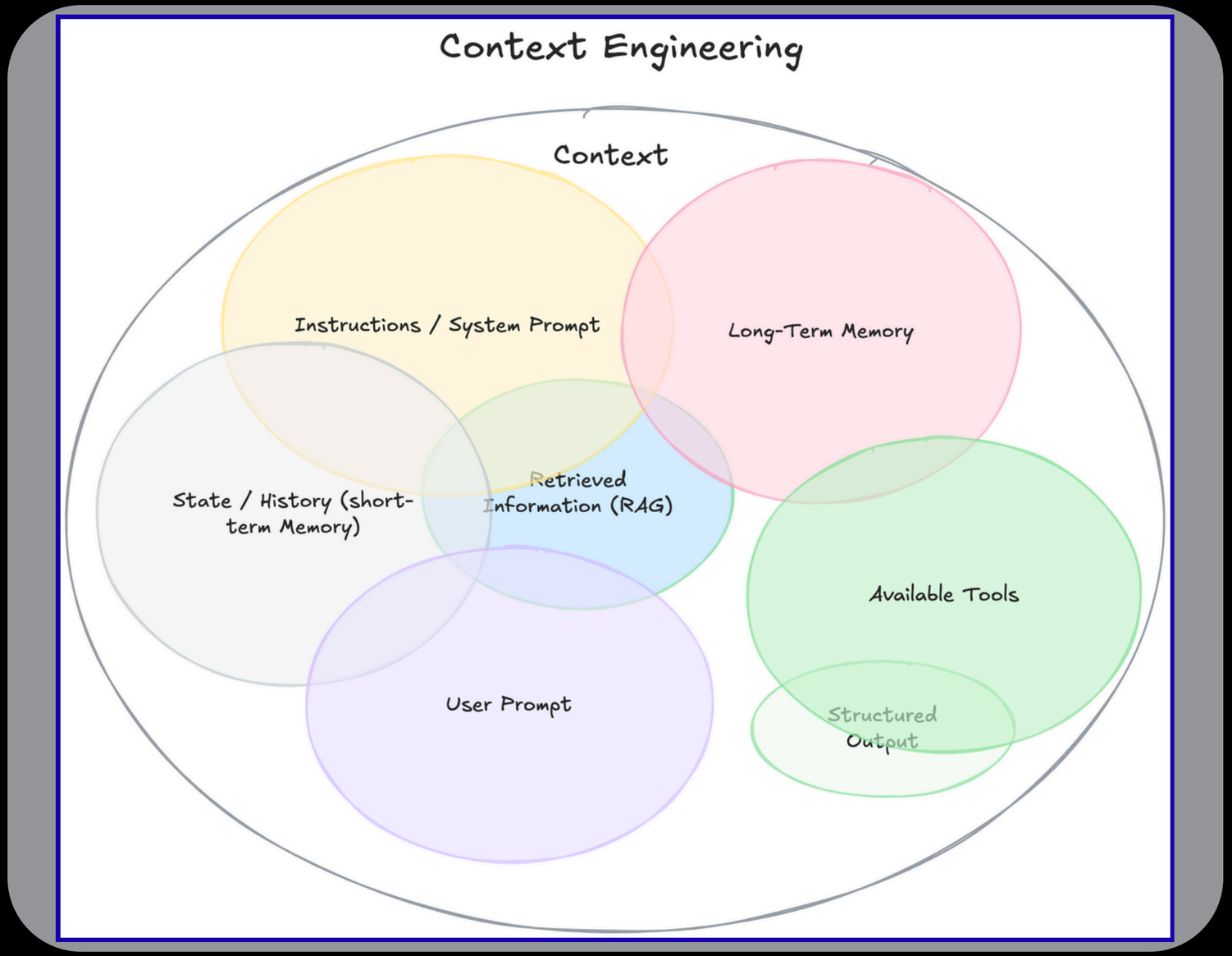


EMBEDDINGS MODELS




1 load documents
2 generate documents chunks
3 vectorize documents chunks
4 store embeddings with document chunk ID
Documents
Document chunks
Embeddings
Vector DB
5 vectorize question
6 use question embedding to retrieve relevant document chunk ID
Embedding
Question
relevant document chunks
7
use document chunk IDs to retrieve document chunks from storage
LLM
Document chunks
8 use question + relevant document chunks + prompt to answer question
9 generated answer
Answer

~~PROMPT~~ CONTEXT ENGINEERING



¿Qué es DOF-RAG?

El Diario Oficial de la Federación (DOF) de México es una fuente masiva de documentos legales y gubernamentales. La información es densa y difícil de consultar.

Leyes, decretos, reglamentos, acuerdos, circulares, resoluciones y demás disposiciones de carácter general que emiten los Poderes de la Unión, órganos constitucionales y otras entidades federales.

Nuestra Misión

Democratizar el acceso a la información legal

Democratizar el conocimiento de cómo funcionan las apps de IA

Pythonistas

¡Únete al Proyecto!

Encuentra nuestro proyecto en GitHub:

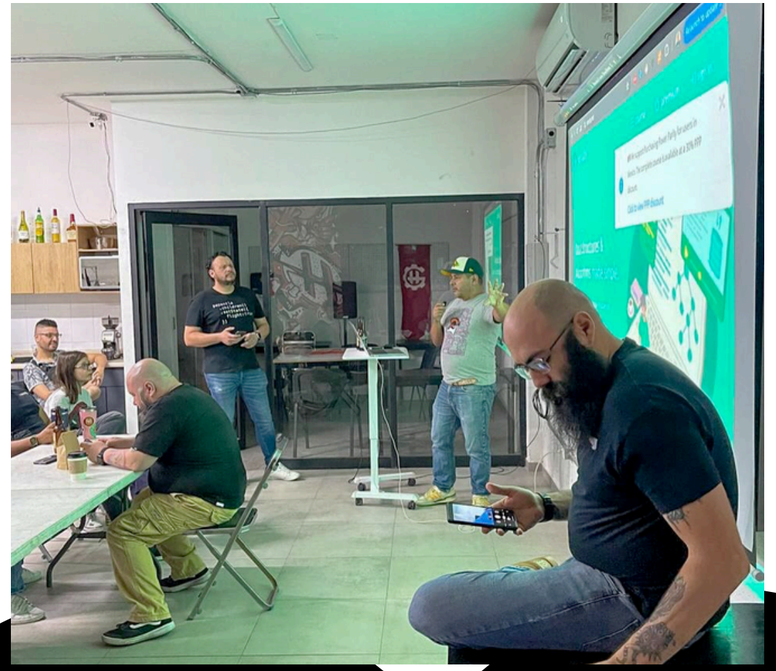
https://github.com/CodeandoGuadalajara/dof-rag

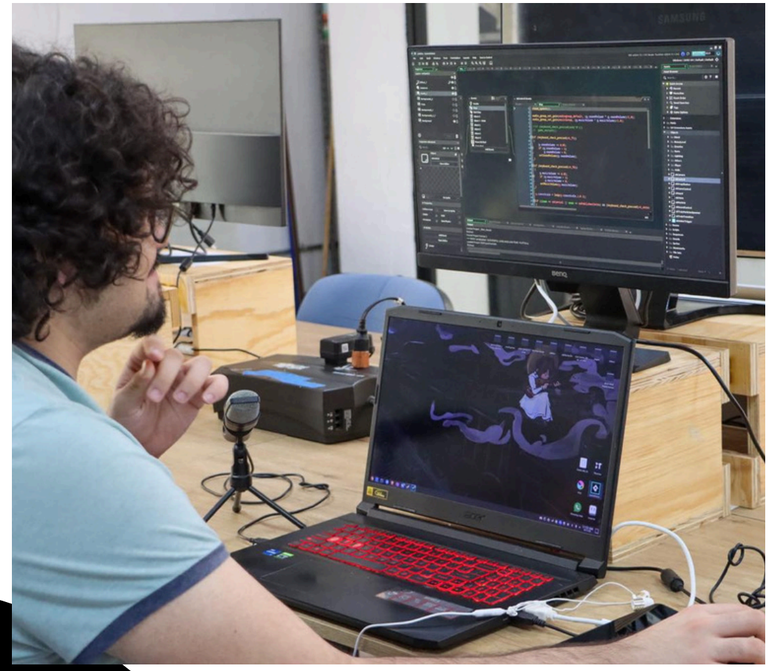
Encuentra nuestro blog donde publicamos las últimas novedades:

https://codeandoguadalajara.github.io/dof-rag-website/es/

HackerGarage

HackerGarage es un espacio único de creación y encuentro para personas interesadas en la tecnología





JÓVENES CONSTRUYENDO EL FUTURO

Jovenes construyendo el futuro

Este es el programa del Gobierno de México de becas para recien egresados para proporcionales experiencia en el mundo laboral. Su objetivo es que sus miembros puedan ser contratados al finalizar o durante el programa de 12 meses.

JOVENES CONSTRUYENDO EL FUTURO

Este programa otorga poco más del salario mínimo de 425 dólares mensuales.

Año	MXN	MXN/USD	USD
2018	$2,651	19.9	$133
2019	$3,080	19.5	$158
2020	$3,697	21	$176
2021	$4,251	20.5	$207
2022	$5,186	20	$259
2023	$6,223	19.9	$313
2024	$7,468	19.8	$377
2025	$8,364	19.7	$425
tech level 1	$34,000	19.7	$1,726
tech senior	$93,800	19.7	$5,000

OTROS RECURSOS

Pythonistas

- Cada quien pone su laptop
- GitHub for non-profit (Codeando México) para el código y la página
- Motherduck.com (duckdb) free de 10GB
- Google colab free tier (jupyter notebook con GPU T4, 1 hora diaria)
- Raspberry Pi v5 (muy lento)
- Considerando speak.com para inglés ($1,000 MXN / $53 USD anuales)
- **Conseguimos** un servidor dedicado Heztner.com de Ryzen 7 PRO, 64 GB RAM, 1 TB SSD, por $53 USD mensual

- **Nuestro tiempo**

Pythonistas

El desafío de la información en imágenes

Gran parte de los datos críticos están en imágenes, invisibles para los sistemas de búsqueda basados en texto

- PDFs diarios de ~300 hojas con cambios a las leyes desde 1920.
- Desde 2006 son PDFs con texto (no fotos)
- El archivo histórico del DOF tiene un aproximado de 25,000 imágenes en 20 años.
- Contienen: tablas de datos, gráficas, infografías, mapas, etc.

Para un sistema de búsqueda tradicional, esta información es invisible

La imagen misteriosa...

Pythonistas

Modelos Multimodales a Prueba

MODELOS LOCALES

MODELO EN LA NUBE (API)

GEMMA3

4.3B PARAMS

MOONDREAM

2B PARAMS

GEMINI-2.0-FLASH

MÁS DE 35B PARAMS

Pythonistas

GEMMA3

PREGUNTA:

¿QUÉ SE OBSERVA EN ESTA IMAGEN?

“LA IMAGEN MUESTRA UNA FOTOGRAFÍA DE UN JOVEN, PRESUMIBLEMENTE UN NIÑO, SENTADO EN UNA SILLA DE RUEDAS. EL NIÑO TIENE UNA EXPRESIÓN SERIA Y SE LE VE CONCENTRADO. LA FOTO PARECE SER DE UN CONTEXTO EDUCATIVO O INFORMATIVO, POSIBLEMENTE RELACIONADA CON LA SENSIBILIZACIÓN SOBRE LA DISCAPACIDAD O LA MOVILIDAD INFANTIL. NO SE IDENTIFICA UN LUGAR O REGIÓN ESPECÍFICA”

Los resultados....

Pythonistas

THE IMAGE PRESENTS A VIBRANT INFOGRAPHIC THAT IS A RIOT OF COLORS AND SHAPES. AT THE HEART OF THE INFOGRAPHIC, A PERSON IS DEPICTED IN A DYNAMIC POSE, SUGGESTING MOVEMENT OR ACTION. THE PERSON'S HEAD IS TILTED TO ONE SIDE, ADDING A DYNAMIC ELEMENT TO THE OTHERWISE STATIC STRUCTURE.

MOONDREAM

PREGUNTA:

¿QUÉ SE OBSERVA EN ESTA IMAGEN?

Los resultados....

Pythonistas

Los resultados....

LA IMAGEN ES UN “PROGRAMA PRECIOS DE GARANTÍA A PRODUCTOS ALIMENTARIOS BÁSICOS: MAÍZ Y FRIJOL”. APARECE EL LOGOTIPO DE SEGALMEX CON EL ESLOGAN “SEGURIDAD ALIMENTARIA MEXICANA”. EL DOCUMENTO ESTÁ ORGANIZADO EN OCHO SECCIONES NUMERADAS QUE EXPLICAN QUE EL PROGRAMA BUSCA INCENTIVAR EL INGRESO DE PEQUEÑOS Y MEDIANOS PRODUCTORES AGROPECUARIOS Y ESTÁ DIRIGIDO A PRODUCTORES DE MAÍZ Y FRIJOL QUE CUMPLAN CON CRITERIOS DE ELEGIBILIDAD.....

GEMINI-2.0-FLASH

PREGUNTA:

¿QUÉ SE OBSERVA EN ESTA IMAGEN?

Modelo	Tiempo Promedio	Limitaciones
Gemini	3-5 segundos	• 15 peticiones/min • ~~1500 peticiones/día~~ • 200 peticiones/dia
gemma3	120-130 segundos	Rendimiento de hardware
Moondream	140-150 segundos	Sólo inglés

Pythonistas

EL PROBLEMA DE LA FALTA DE CONTEXTO

Berlin[a] is the capital and largest city of Germany, both by area and by population.[11] Its more than 3.85 million inhabitants[12] make it the European Union's most populous city, as measured by population within city limits.[13] The city is also one of the states of Germany, and is the third smallest state in the country in terms of area. Berlin is surrounded by the state of Brandenburg, and Brandenburg's capital Potsdam is nearby. The

`https://en.wikipedia.org/wiki/Berlin`

Full Wikipedia Article

- "Berlin is the capital and largest city of Germany, both by area and by population."
- "Its more than 3.85 million inhabitants make it the European Union's most populous city, as measured by population within city limits."
- "The city is also one of the states of Germany, and is the third smallest state in the country in terms of area."

Should be derived from context (other chunks)

Chunked into Sentences

El problema de la falta de contexto

Nike's 2023 10-K headings

Form 10-K Cover Page and Company Information
Table of Contents
Business Overview
Product Offerings
Sales and Marketing Strategy
Market Segments and Geographic Operations
Significant Customer Information
Product Research, Design and Development
Manufacturing and Supply Chain Overview
International Operations and Trade Challenges

Document: Nike 10-K FY2023
Section: Operating Segments Overview

← Chunk header

Asia Pacific & Latin America(2) 6,431 5,955 8 % 17 % 5,343 11 % 16 %
Global Brand Divisions(3) 58 102 -43 % -43 % 25 308 % 302 %
TOTAL NIKE BRAND $ 48,763 $ 44,436 10 % 16 % $ 42,293 5 % 6 %
Converse 2,427 2,346 3 % 8 % 2,205 6 % 7 %
Corporate(4) 27 (72) – – 40 – –
TOTAL NIKE, INC. REVENUES $ 51,217 $ 46,710 10 % 16 % $ 44,538 5 % 6 %
(1) The percent change excluding currency changes represents a non-GAAP financial measure. For further information, see "Use of Non-GAAP Financial Measures".
(2) For additional information on the transition of our NIKE Brand businesses within our CASA territory to a third-party distributor, see Note 18 –

Query: Nike operating segment results

Similarity with contextual chunk header: 0.9479
Similarity without contextual chunk header: 0.7077

Pythonistas

Enfoque (Estructurado)

```
# DOCUMENT: 01052025-DOF
## TÍTULO 1
### SUBTÍTULO A
#### APARTADO 3
```

Pythonistas

El Embedding Perfecto

Con los datos extraídos de las imágenes y contextualizados en secciones, enfrentamos el siguiente reto: elegir el motor que los convertirá en vectores

Modelos Evualuados

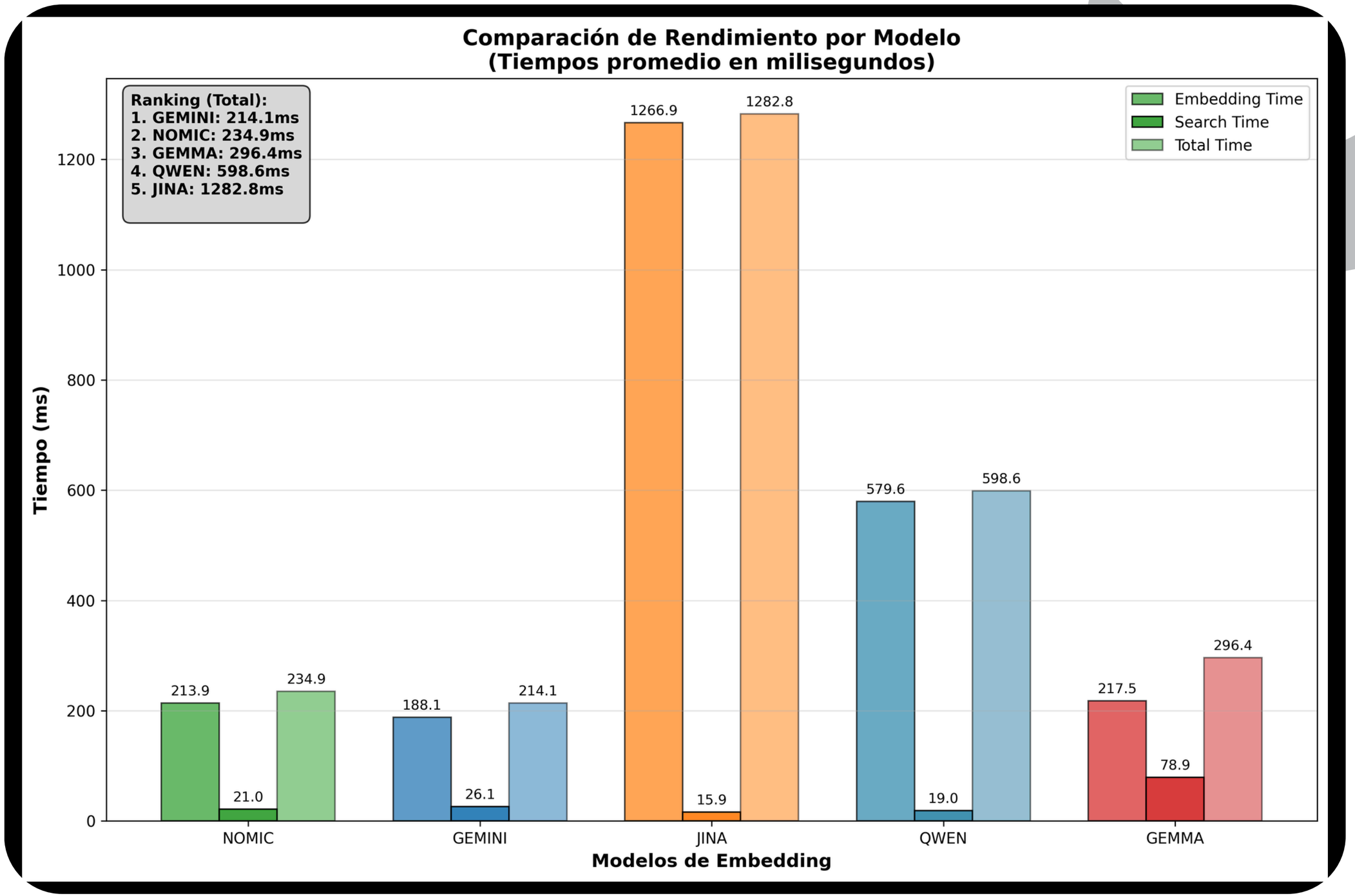
- GEMINI-EMBEDDING
- NOMIC-AI/MODERNBERT-EMBED-BASE
- JINA-EMBEDDINGS-V4
- GEMMA EMBEDDING 300M
- QWEN3 EMBEDDING 0.6B

Metricas

Comparación

Medimos el tiempo desde que llega una consulta hasta que se entregan los resultados. Esto incluye generar el embedding de la pregunta y buscar en la base de datos

Rendimiento Total (Embedding + Búsqueda)


Comparación de Rendimiento por Modelo
(Tiempos promedio en milisegundos)
Ranking (Total):
1. GEMINI: 214.1ms
2. NOMIC: 234.9ms
3. GEMMA: 296.4ms
4. QWEN: 598.6ms
5. JINA: 1282.8ms
Embedding Time
Search Time
Total Time
Tiempo (ms)
0
200
400
600
800
1000
1200
213.9
21.0
234.9
188.1
26.1
214.1
1266.9
15.9
1282.8
579.6
19.0
598.6
217.5
78.9
296.4
NOMIC
GEMINI
JINA
QWEN
GEMMA
Modelos de Embedding

EN CONSULTA..

¿CUÁL ES EL HORARIO DE IMPORTACIÓN Y EXPORTACIÓN DE LA ADUANA DE AGUASCALIENTES?

NOMIC
(TITLE: 06012025-MAT)

- PÁGINA 9
- PÁGINA 7
- PÁGINA 6
- PÁGINA 10
- PÁGINA 8

GEMINI
(TITLE: 06012025-MAT)

- **PÁGINA 5**
- PÁGINA 7
- PÁGINA 8
- PÁGINA 9
- PÁGINA 6

JINA
(TITLE: 06012025-MAT)

- **PÁGINA 5**
- PÁGINA 6
- PÁGINA 7
- PÁGINA 8
- PÁGINA 10

EN CONSULTA..

¿CUÁL ES EL HORARIO DE IMPORTACIÓN Y EXPORTACIÓN DE LA ADUANA DE AGUASCALIENTES?

QWEN

(TITLE: 06012025-MAT)

- **PÁGINA 5**
- PÁGINA 7
- PÁGINA 9
- PÁGINA 10
- PÁGINA 8

GEMMA

(TITLE: 06012025-MAT)

- PÁGINA 8
- PÁGINA 10
  (TITLE: 09012025-MAT)
- PÁGINA 122
  (TITLE: 17012025-MAT)
- PÁGINA 14

Por lo que....

GEMINI

- **CALIDAD:** MUY ALTA, RESULTADOS SUPERIORES.
- **VELOCIDAD:** SUMAMENTE RÁPIDO.
- **ACCESO:** EXTERNO, TOTAL DEPENDENCIA DE LA API.
- **COSTES:** SUJETO A PRECIOS Y LÍMITES DE USO.
- **CONTROL:** NULO SOBRE CAMBIOS Y VERSIONES FUTURAS.

NOMIC

- **CALIDAD:** FUNCIONAL, AUNQUE NO TAN REFINADA.
- **VELOCIDAD:** RÁPIDO POR SU TAMAÑO REDUCIDO.
- **RECURSOS:** LIGERO Y ACCESIBLE.
- **CONTROL:** TOTALMENTE LOCAL, GARANTIZANDO MÁXIMA PRIVACIDAD.

Por lo que....

QWEN

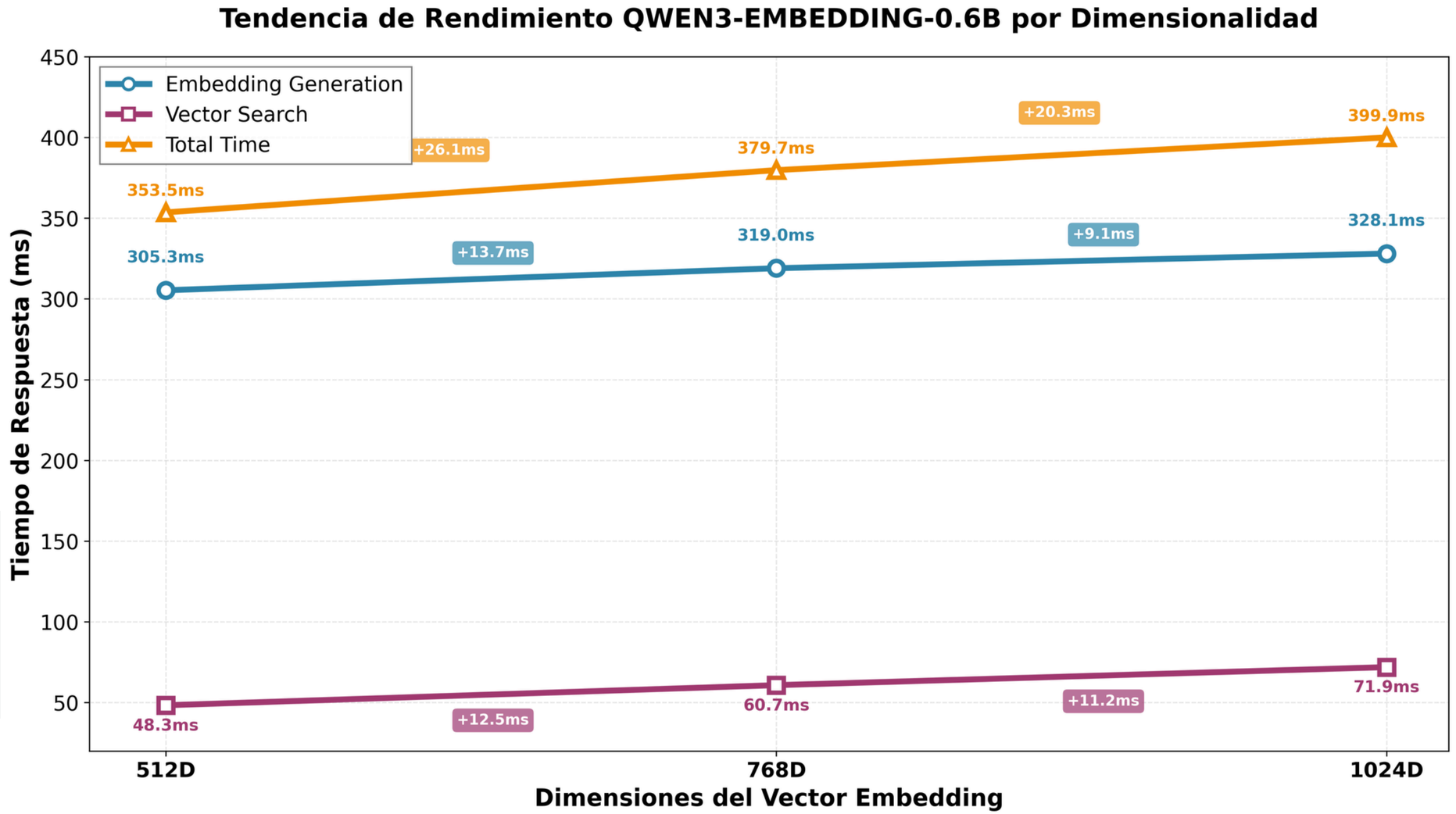
- **CALIDAD:** ALTA Y CONSISTENTE.
- **VELOCIDAD:** LENTO Y ESTABLE.
- **RECURSOS:** EXIGENTE (REQUIERE HARDWARE ESPECÍFICO).
- **CONTROL:** TOTALMENTE LOCAL, GARANTIZANDO MÁXIMA PRIVACIDAD.

GEMMA

- **CALIDAD:** VARIABLE (REQUIERE UN MAYOR TRATAMIENTO DE LOS DATOS).
- **VELOCIDAD:** MUY RÁPIDO.
- **RECURSOS:** LIGERO Y DE BAJO CONSUMO.
- **LIMITACIÓN:** ENTRADA DE TEXTO DE SOLO 2048 TOKENS.
- **CONTROL:** TOTALMENTE LOCAL, GARANTIZANDO MÁXIMA PRIVACIDAD.

Pythonistas

Velocidad de embedding por dimensión Qwen.



Pythonistas

Velocidad de embedding por dimensión Qwen.

Dimensión	Embedding (ms)	Búsqueda (ms)	Total (ms)	Consultas/ seg	Incremento vs 512
512D	305.3	48.3	353.5	2.83	Base
768D	319.0	60.7	379.7	2.63	+7.4%
1024D	328.1	71.9	399.9	2.50	+13.1%

Estimaciones de Almacenamiento: Base de Datos DuckDB

Las siguientes proyecciones se refieren exclusivamente al espacio requerido por la base de datos DuckDB para el proyecto DOF-RAG.

Más detalles en nuestro post

Objetivo: Decidir el tamaño ideal para los "embeddings" (vectores de datos) sin sacrificar rendimiento ni gastar de más en almacenamiento a futuro.

- ¿Qué se analizó?
  - Se compararon tres dimensiones de embeddings: 512d, 768d y 1024d.
  - Se usó una muestra real de 10,090 fragmentos de documentos de enero de 2025.
- Hallazgo Principal:
  - La base de datos (DuckDB) siempre ocupa casi 3 veces el tamaño de los datos crudos, sin importar la dimensión. Este costo extra es para optimizar la velocidad y es un factor constante.

Estimaciones de Almacenamiento: Base de Datos DuckDB

La Decisión y Las Cifras Finales

La Decisión: Se eligió la dimensión de 768d para el proyecto.

- ¿Por qué 768d?
  - Balance Perfecto: Ofrece la misma calidad de respuesta que la opción más grande (1024d).
  - Eficiencia: Ahorra casi un 10% de espacio de almacenamiento a largo plazo.
  - Sostenibilidad: Es la opción más equilibrada para crecer durante 25 años.

- Proyecciones Oficiales de Almacenamiento:
  - A 1 año: Se necesitarán 2.44 GB.
  - A 25 años: Se estima un total de ~60 GB.

Más detalles en nuestro post

Pythonistas

Cambio de Metodología: De Páginas a Tokens

Razón 1: Prioridad en la Calidad de Datos

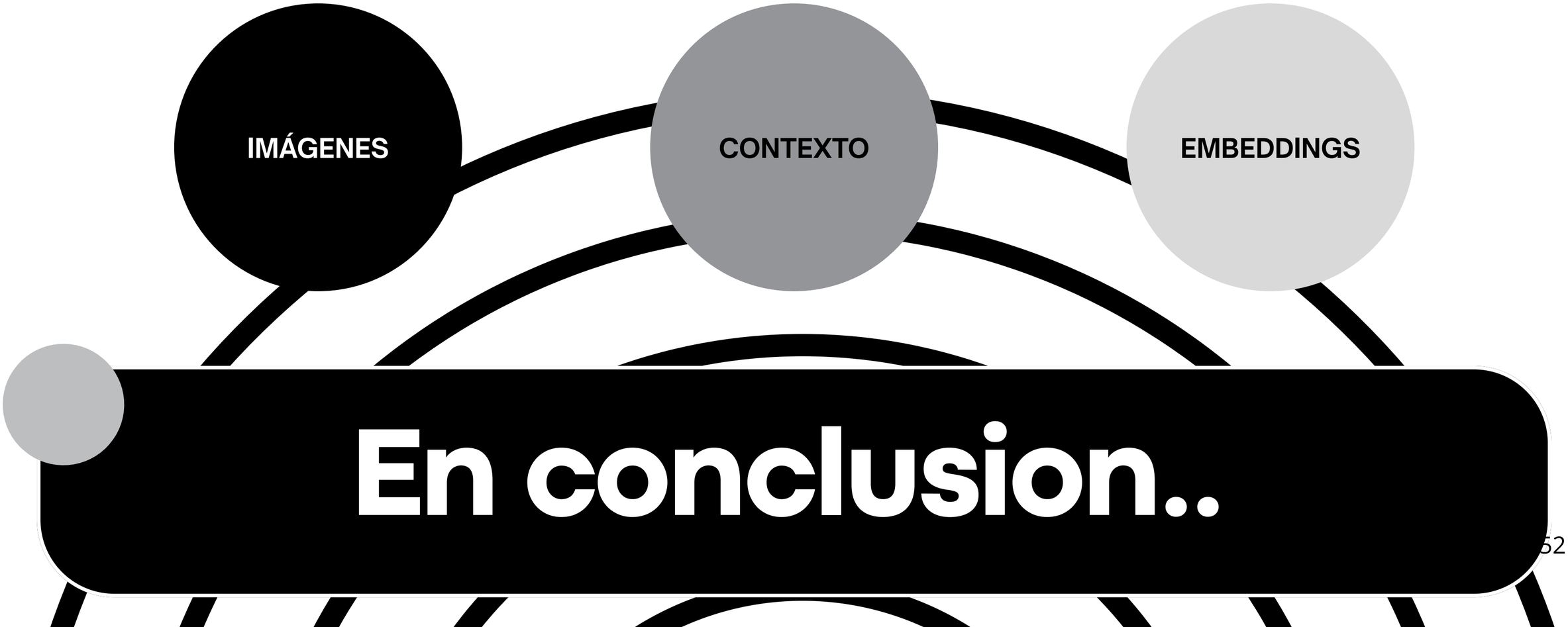
- **Extracción Superior:** El nuevo flujo con Pandoc y Word (DOCX) garantiza tablas y encabezados sin errores.
- **Decisión Estratégica:** Se priorizó la integridad de los datos sobre la paginación para asegurar un proceso más robusto y evitar dependencias técnicas complejas (MS Word en Windows).

Cambio de Metodología: De Páginas a Tokens

Razón 2: Adaptación al Modelo Gemma

- **Requisitos del Modelo:** Gemma es más rápido y eficiente, pero exige datos limpios y tiene un límite estricto de 2048 tokens.
  - **Nuestra Solución:** Se ajustaron los fragmentos (chunks) a 1024 tokens para operar de forma segura dentro del límite, maximizando el rendimiento y la calidad.

Construir un sistema RAG es un viaje de retos interconectados en extracción de datos (imágenes), preservación de contexto (contexto) y selección de modelos (embeddings).



En conclusion..

SIGUIENTES RETOS

PUBLICAR LA INTERFAZ CHAT WEB

EXTRAER INFORMACIÓN Y EMBEDDINGS DE 2006 A LA FECHA (~40GB)

LO MISMO, PERO DESDE 1920??? :-P (~212GB?)

MEJORAR LA EXTRACCIÓN DE INFORMACIÓN DE TABLAS

CONSIDERAR CAMBIAR DE ARCHIVOS PDF A DOCX (MEJOR CALIDAD DE TABLAS)

MEJORAR LAS PRUEBAS DE EVALUACIÓN (MÁS PREGUNTAS)

MEJORAR LOS MÉTODOS DE BÚSQUEDA

Pythonistas

¿Preguntas?

Pythonistas

Gracias!