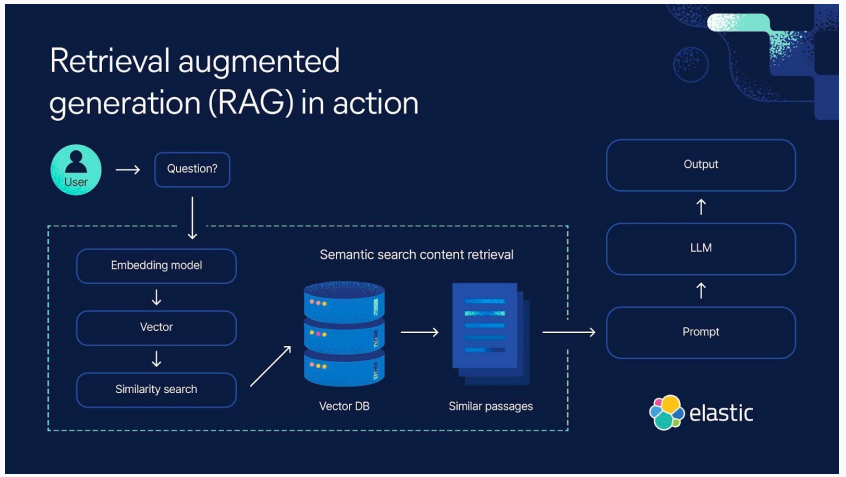
THE BRIDGE

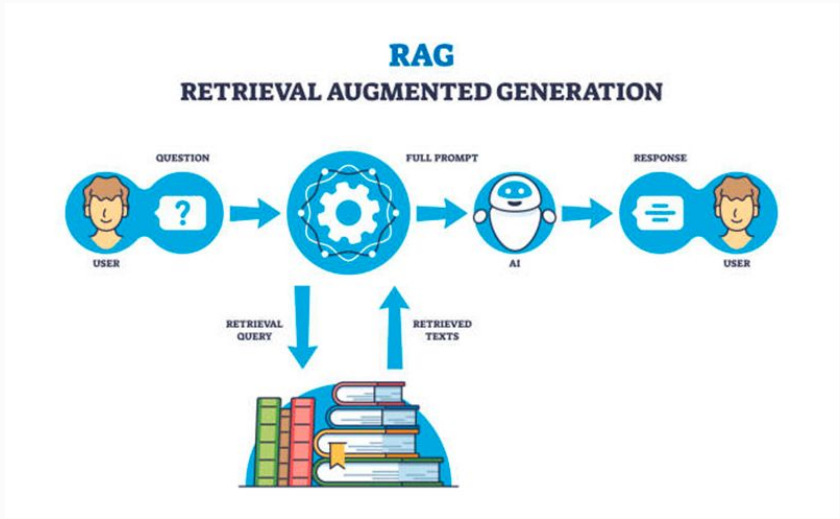
Fundamentos de los sistemas RAG

Introduction

1. Introducción: ¿Qué es un sistema RAG?

- **RAG** significa *Retrieval-Augmented Generation*
- Es una **arquitectura de IA** que combina dos cosas:
    - **Recuperación de información (retrieval)** de fuentes externas.
    - **Generación de texto (generation)** con un modelo de lenguaje (como GPT, LLaMA, etc.).
- Su objetivo: **mejorar la precisión y actualidad** de las respuestas del modelo, usando información externa en tiempo real.





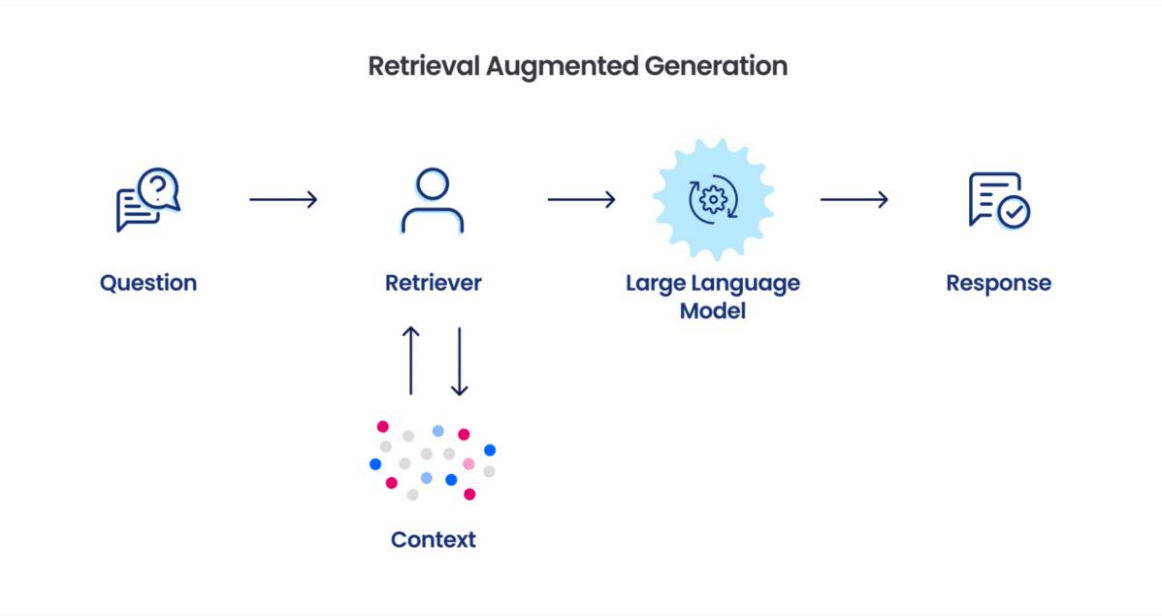
- Ejemplo:

  Si le preguntas a un modelo RAG “¿Cuáles son los últimos descubrimientos sobre energía solar?”, el sistema busca en una base de datos o internet documentos recientes, **recupera los más relevantes**, y el generador produce una **respuesta en lenguaje natural basada en esa información**.

2. Componentes principales de un sistema RAG

a) Retriever (Recuperador)

- Es la parte que **busca información relevante** en una base de conocimiento (documentos, artículos, bases de datos, etc.).
- Usa **embeddings** (vectores numéricos que representan significado) para encontrar textos **semánticamente similares** a la pregunta.
- Técnicas comunes: **búsqueda semántica**, **vector databases** (como FAISS, Pinecone, Milvus).



🔹 Ejemplo:
Si el usuario pregunta *"¿Cómo funciona una célula solar?"*, el retriever busca los documentos más parecidos en significado, no solo en palabras exactas.

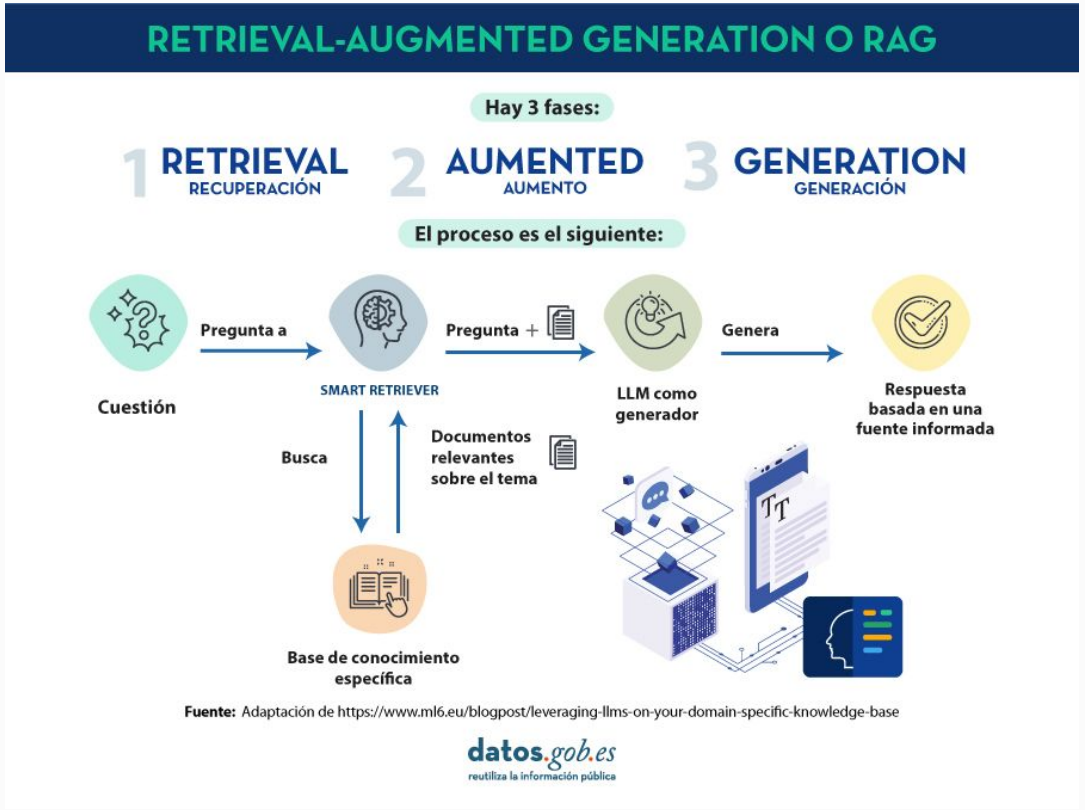
b) Generator (Generador)

- Es el **modelo de lenguaje** (como GPT o T5) que **lee la información recuperada** y **genera una respuesta coherente y natural**.

- Integra los fragmentos encontrados y los usa como contexto para redactar una respuesta completa.

🔹 En otras palabras:

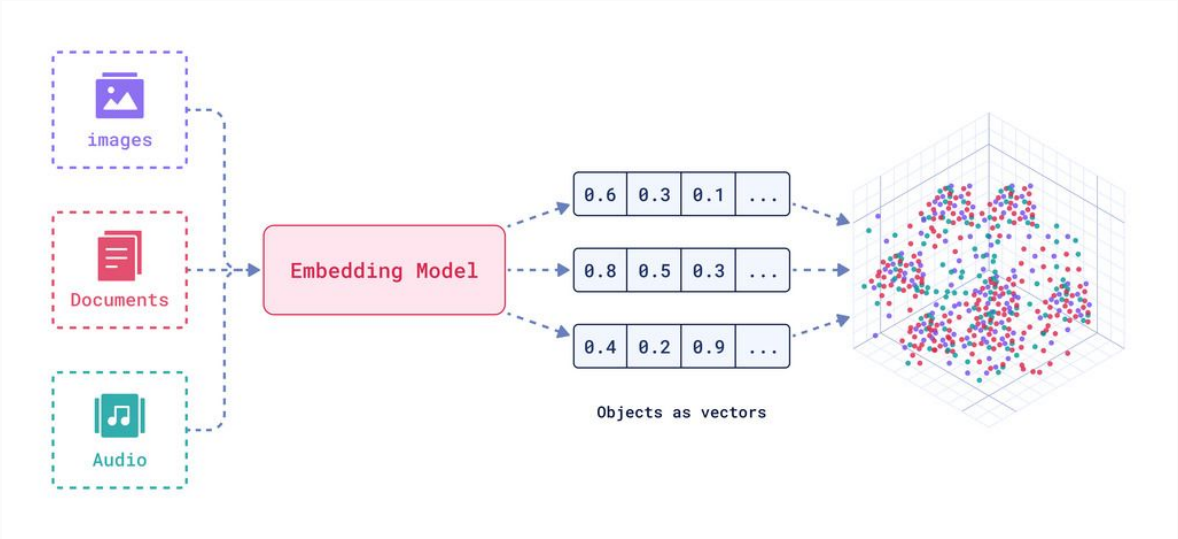
El retriever encuentra, el generador explica.

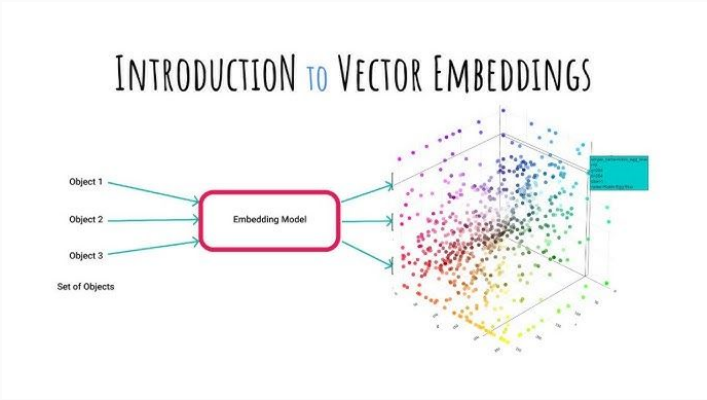


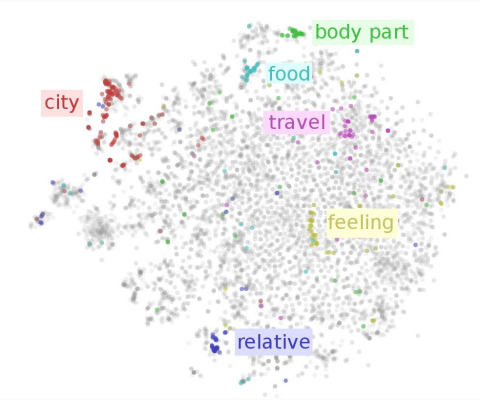
3. Importancia de los *embeddings* y la búsqueda semántica

Embeddings

- Son **representaciones numéricas** de textos, creadas por modelos de IA.
- Permiten medir la **similitud de significado** entre frases, aunque no usen las mismas palabras.
- Ejemplo:
  - “auto” y “vehículo” tendrán embeddings muy parecidos.
  - “auto” y “banana” estarán muy lejos en el espacio vectorial.



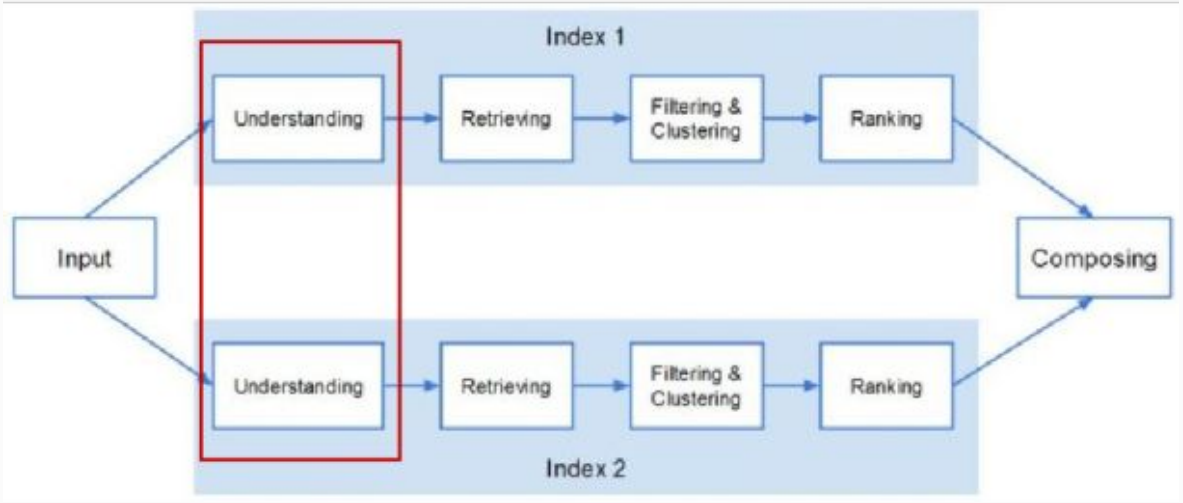


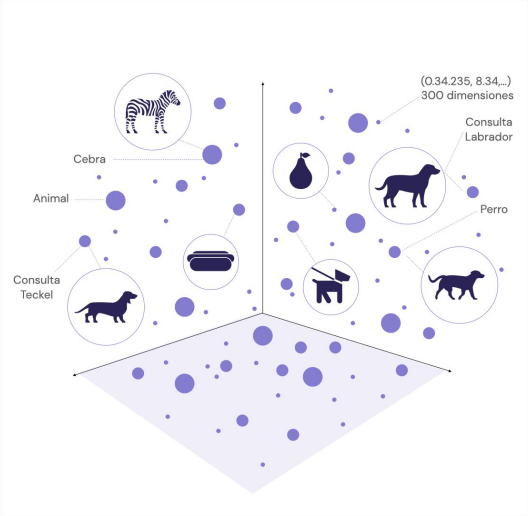


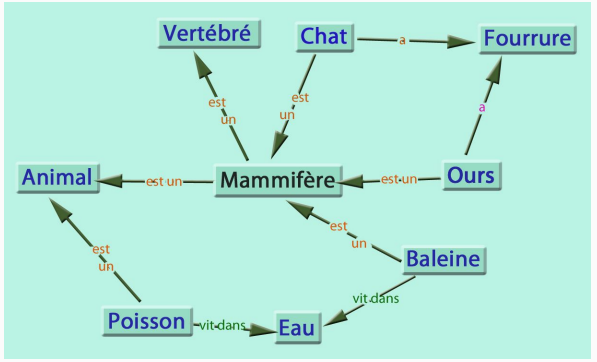
Búsqueda semántica

- Usa esos embeddings para encontrar documentos **por significado**, no por coincidencia exacta de palabras.

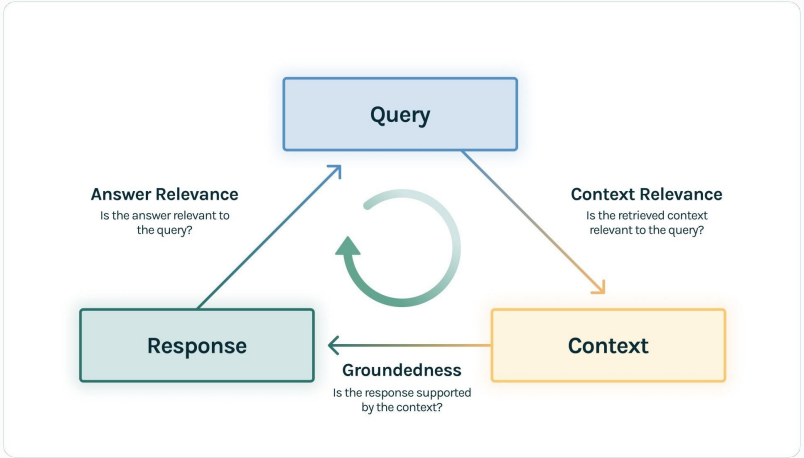
- Es clave en RAG porque permite que el sistema encuentre información **relevante aunque esté redactada de forma distinta**.







4. Por qué los RAG son importantes

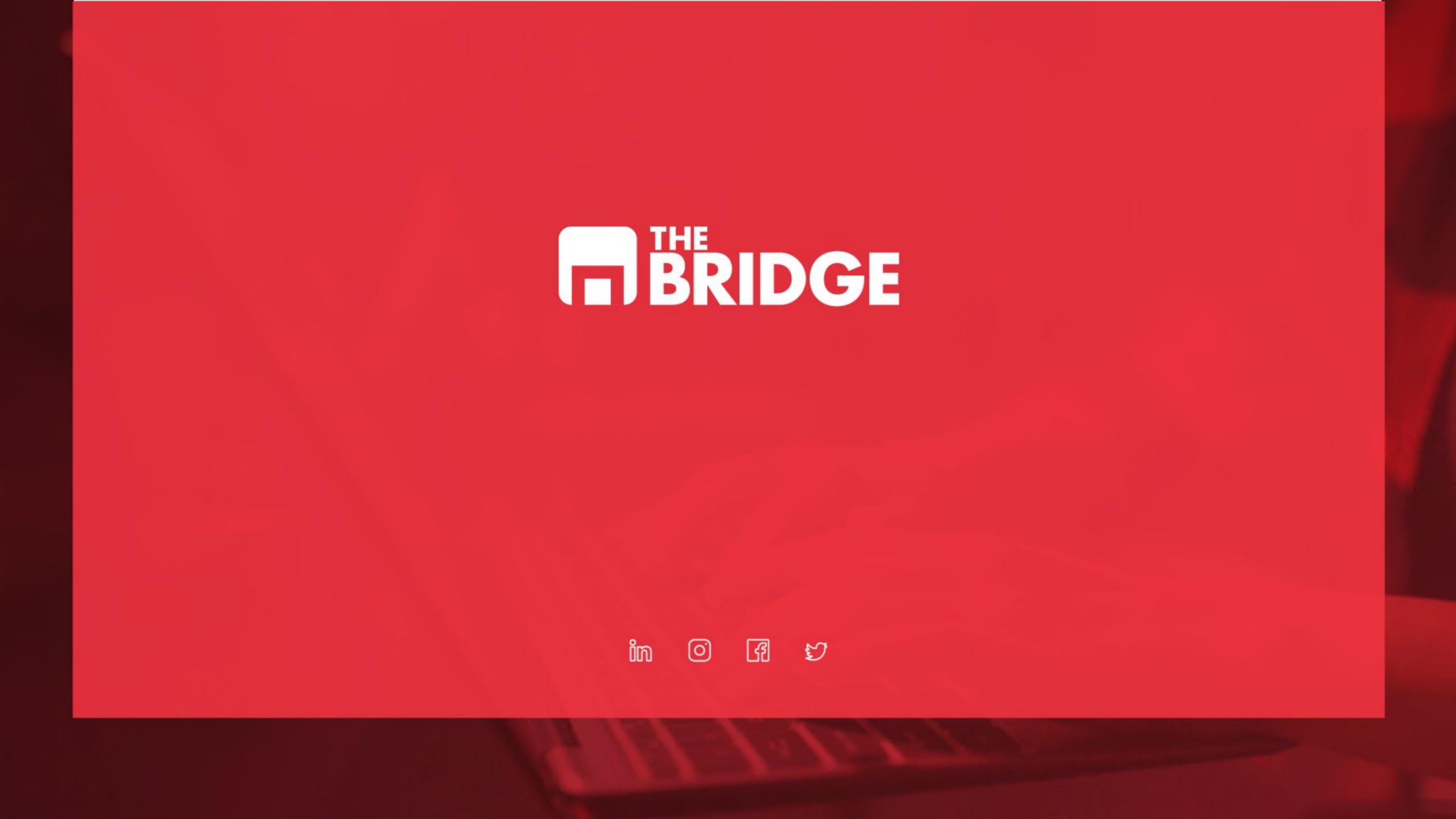


Ventajas:

1. **Actualización continua**: puedes cambiar o ampliar la base de conocimiento sin volver a entrenar el modelo.
2. **Menos errores y alucinaciones**: el modelo se apoya en datos verificables.
3. **Personalización**: puedes crear un RAG específico para tu empresa, universidad o área.
4. **Eficiencia de costos**: entrenar un LLM cuesta millones, pero un RAG usa modelos existentes con bases de datos externas.

Ejemplos de uso avanzado:

- **ChatGPT con "contexto de empresa"** (retrieval de documentos internos).
- **Sistemas legales** que buscan jurisprudencia actual.
- **Asistentes científicos** que leen papers en tiempo real (por ejemplo, Semantic Scholar + RAG).
- **Atención al cliente** que usa manuales y FAQs para responder con precisión.


THE
BRIDGE