

Casos de Uso Específicos y Desafíos Futuros

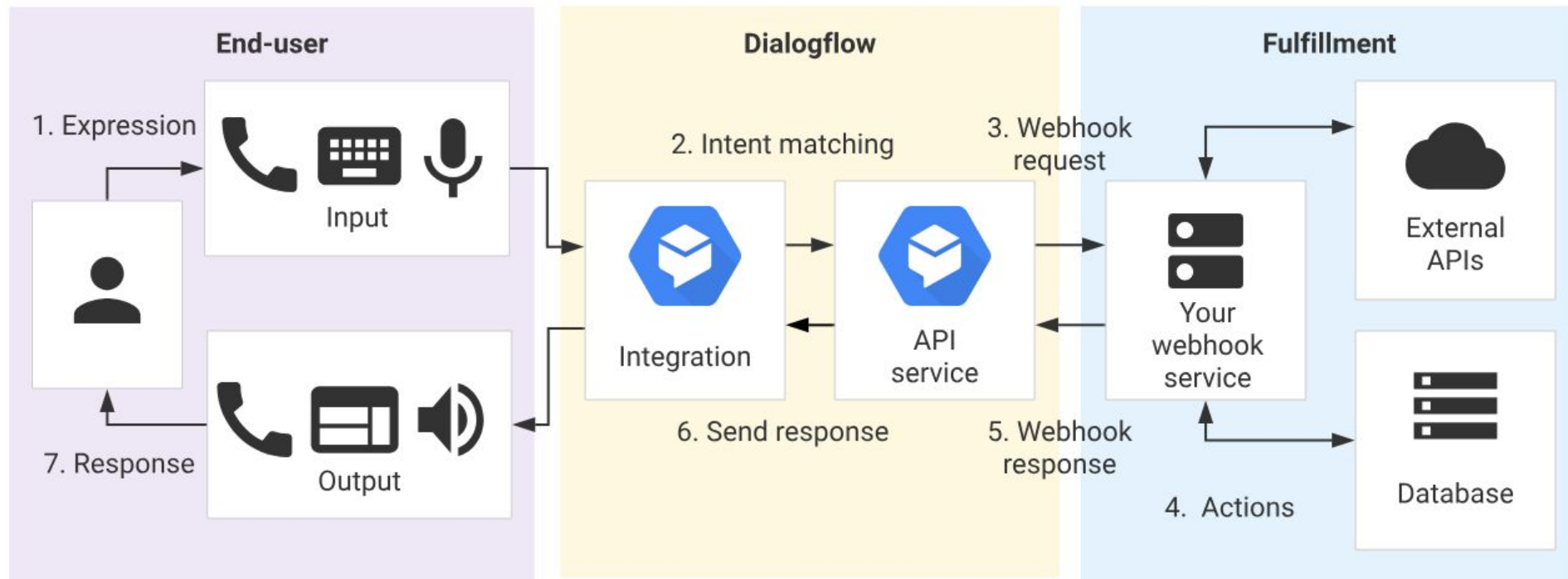
OBS Business School

Introducción

- Agentes conversacionales
- Information Retrieval
- Retrieval-Augmented Generation
- Casos de éxito
- Desafíos futuros

Agentes conversacionales

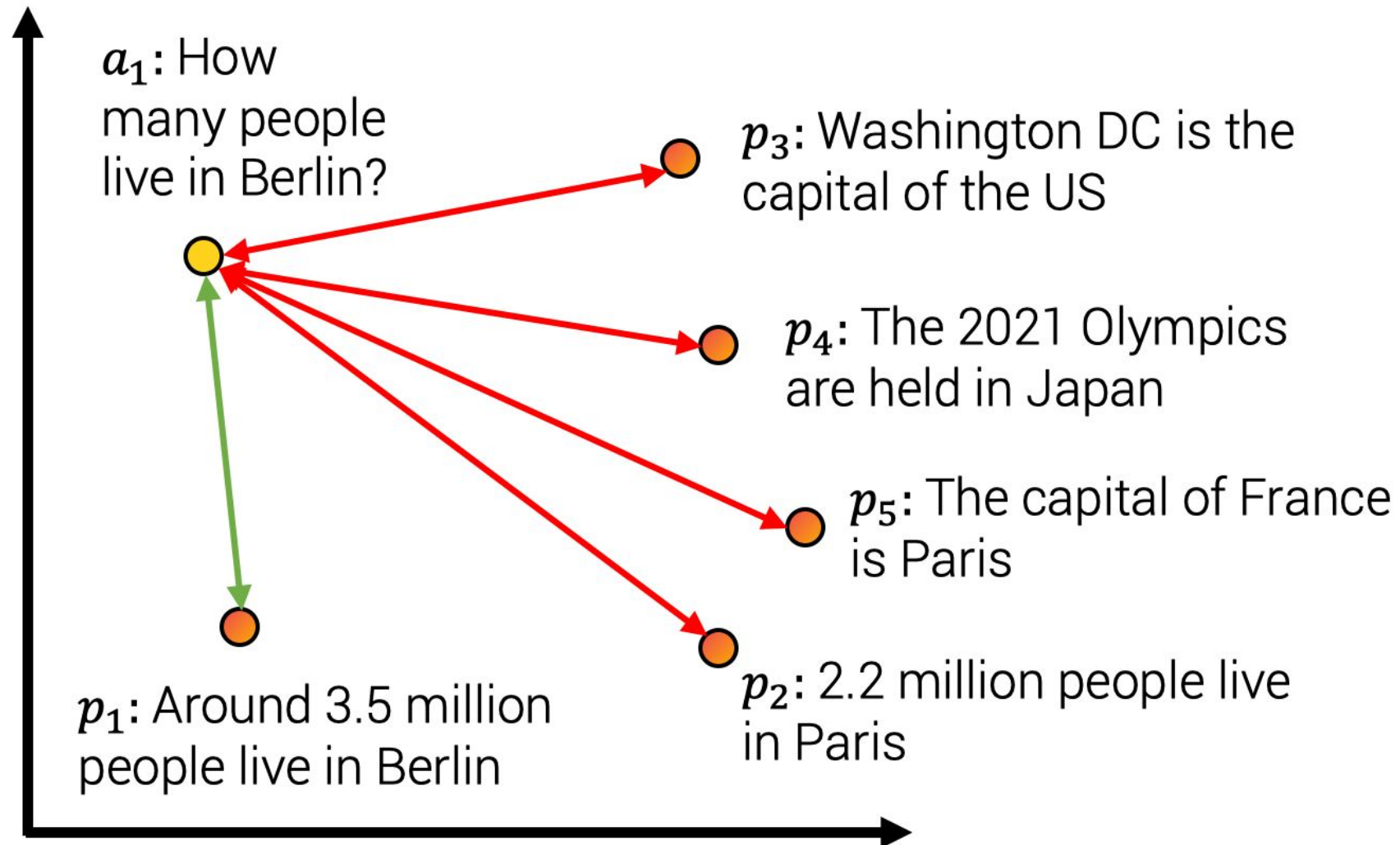
Un agente conversacional es un sistema diseñado para interactuar con los usuarios a través del lenguaje natural. Este tipo de agente puede realizar tareas, responder preguntas y mantener conversaciones de manera similar a un ser humano.



Componentes de un Agente Conversacional

- Interfaz de usuario:
 - Entrada: Permite a los usuarios ingresar texto o voz.
 - Salida: Muestra las respuestas del agente en texto o las reproduce en voz
- Procesamiento de Lenguaje Natural (NLP):
 - Análisis: Comprende y analiza el lenguaje natural del usuario.
 - Generación: Produce respuestas en lenguaje natural.
- Gestión del diálogo:
 - Control del Flujo: Gestiona el flujo de la conversación, asegurando coherencia y contexto.
 - Manejo de Contexto: Mantiene la memoria del contexto de la conversación para referencias futuras.
- Base de conocimiento:
 - Datos: Información y datos almacenados que el agente usa para responder preguntas.
 - Modelos: Algoritmos y modelos que ayudan a interpretar preguntas y generar respuestas.
- Módulo de Integración:
 - APIs: Conecta con otros sistemas y servicios para obtener información o realizar acciones.
 - Servicios externos: Integra con bases de datos, aplicaciones empresariales y otros servicios necesarios.

Information Retrieval



La librería **Sentence Transformer** es una herramienta de que facilita la transformación de oraciones y textos en vectores de alta dimensión (embeddings). Estos vectores representan el significado semántico de las oraciones, permitiendo compararlas y analizarlas de manera eficiente.

¿Para qué sirve?

- **Búsqueda Semántica:** Permite buscar documentos o textos relevantes basándose en el significado, en lugar de palabras clave exactas.
- **Clasificación de textos:** Facilita la categorización de textos en diferentes clases o temas basándose en su contenido.
- **Clustering:** Agrupa textos similares en clusters, lo cual es útil para organizar grandes conjuntos de datos.
- **Reducción de dimensionalidad:** Convierte textos largos en representaciones más compactas que aún retienen el significado.
- **Detección de duplicados:** Identifica textos o documentos que son similares en contenido, ayudando a eliminar redundancias.

```
from sentence_transformers import SentenceTransformer

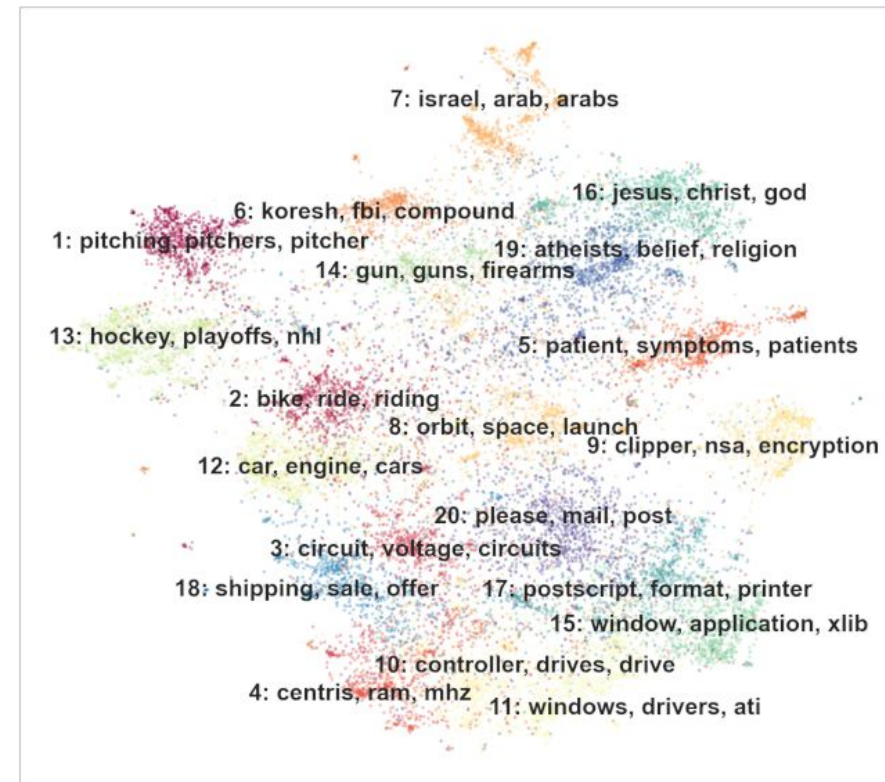
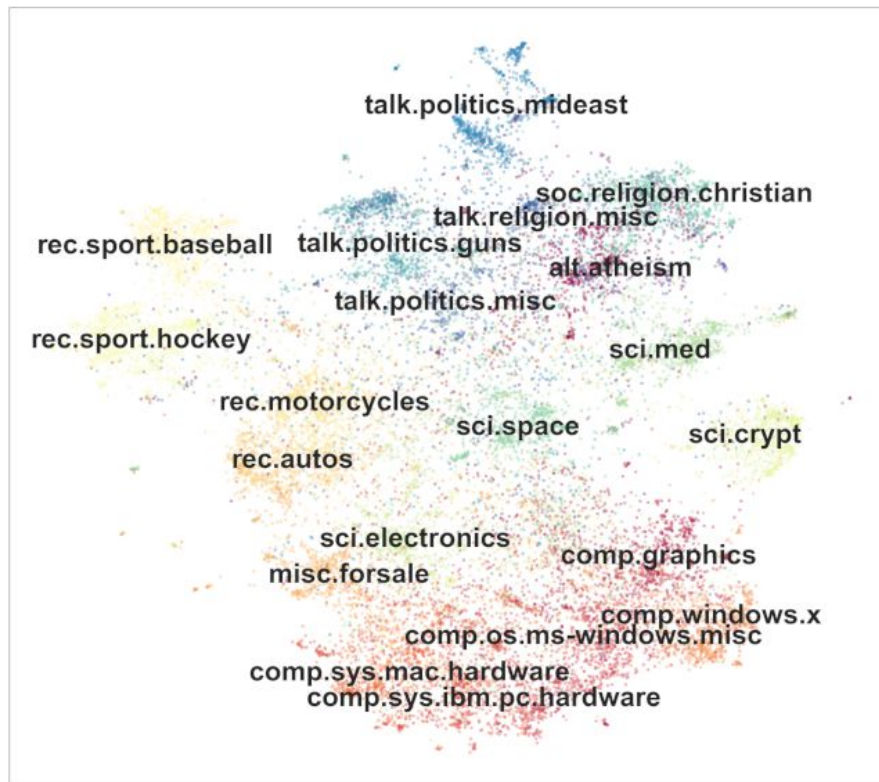
# 1. Load a pretrained Sentence Transformer model
model = SentenceTransformer("all-MiniLM-L6-v2")

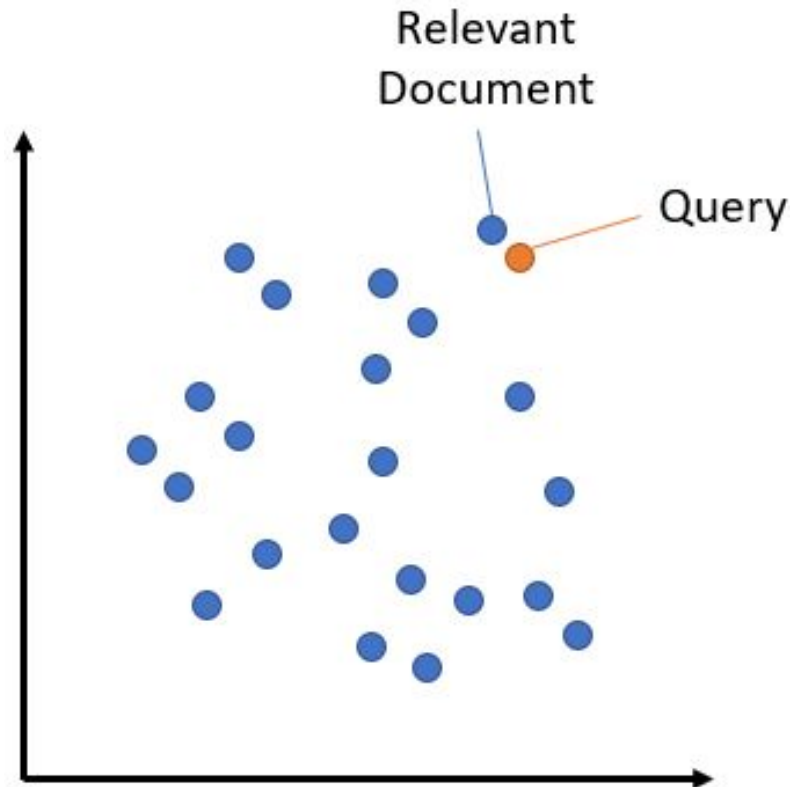
# The sentences to encode
sentences = [
    "The weather is lovely today.",
    "It's so sunny outside!",
    "He drove to the stadium.",
]

# 2. Calculate embeddings by calling model.encode()
embeddings = model.encode(sentences)
print(embeddings.shape)
# [3, 384]

# 3. Calculate the embedding similarities
similarities = model.similarity(embeddings, embeddings)
print(similarities)
# tensor([[1.0000, 0.6660, 0.1046],
#         [0.6660, 1.0000, 0.1411],
#         [0.1046, 0.1411, 1.0000]])
```

- El Topic Modeling es una técnica de procesamiento de lenguaje natural utilizada para identificar temas ocultos dentro de un conjunto de textos.
- Estos temas son conjuntos de palabras que frecuentemente aparecen juntas y representan un concepto o topic.
- Tutorial con Sentence Transformer: <https://www.sbert.net/examples/applications/clustering/README.html>





```
import torch

from sentence_transformers import SentenceTransformer

embedder = SentenceTransformer("all-MiniLM-L6-v2")

# Corpus with example sentences
corpus = [
    "A man is eating food.",
    "A man is eating a piece of bread.",
    "The girl is carrying a baby.",
    "A man is riding a horse.",
    "A woman is playing violin.",
    "Two men pushed carts through the woods.",
    "A man is riding a white horse on an enclosed ground.",
    "A monkey is playing drums.",
    "A cheetah is running behind its prey.",
]

# Use "convert_to_tensor=True" to keep the tensors on GPU (if available)
corpus_embeddings = embedder.encode(corpus, convert_to_tensor=True)

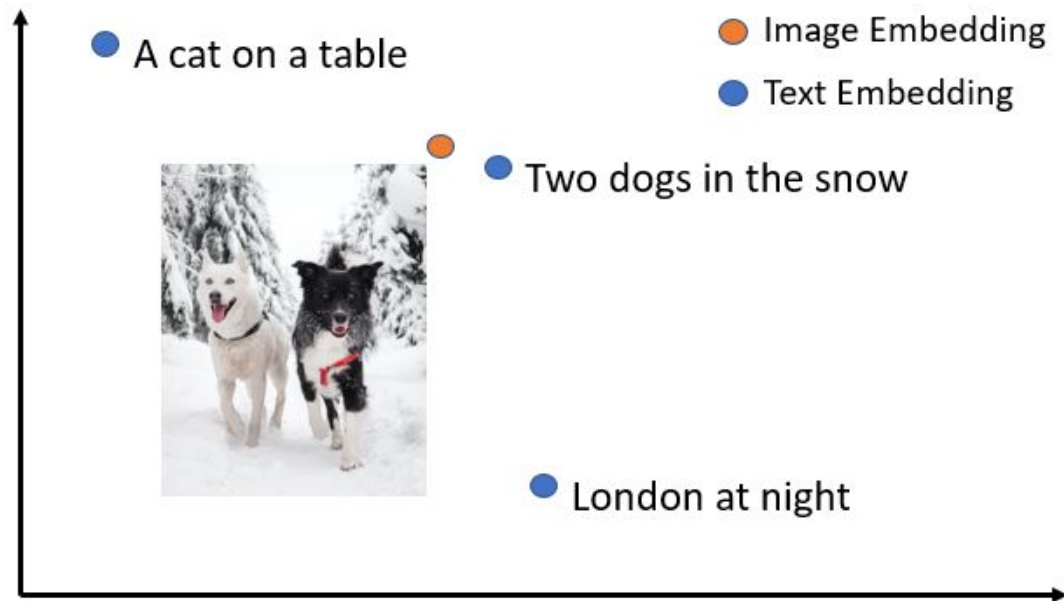
# Query sentences:
queries = [
    "A man is eating pasta.",
    "Someone in a gorilla costume is playing a set of drums.",
    "A cheetah chases prey on across a field.",
]

# Find the closest 5 sentences of the corpus for each query sentence based on cosine similarity
top_k = min(5, len(corpus))
for query in queries:
    query_embedding = embedder.encode(query, convert_to_tensor=True)

    # We use cosine-similarity and torch.topk to find the highest 5 scores
    similarity_scores = embedder.similarity(query_embedding, corpus_embeddings)[0]
    scores, indices = torch.topk(similarity_scores, k=top_k)

    print("\nQuery:", query)
    print("Top 5 most similar sentences in corpus:")

    for score, idx in zip(scores, indices):
        print(corpus[idx], "(Score: {:.4f})".format(score))
```



```
from sentence_transformers import SentenceTransformer
from PIL import Image

# Load CLIP model
model = SentenceTransformer("clip-ViT-B-32")

# Encode an image:
img_emb = model.encode(Image.open("two_dogs_in_snow.jpg"))

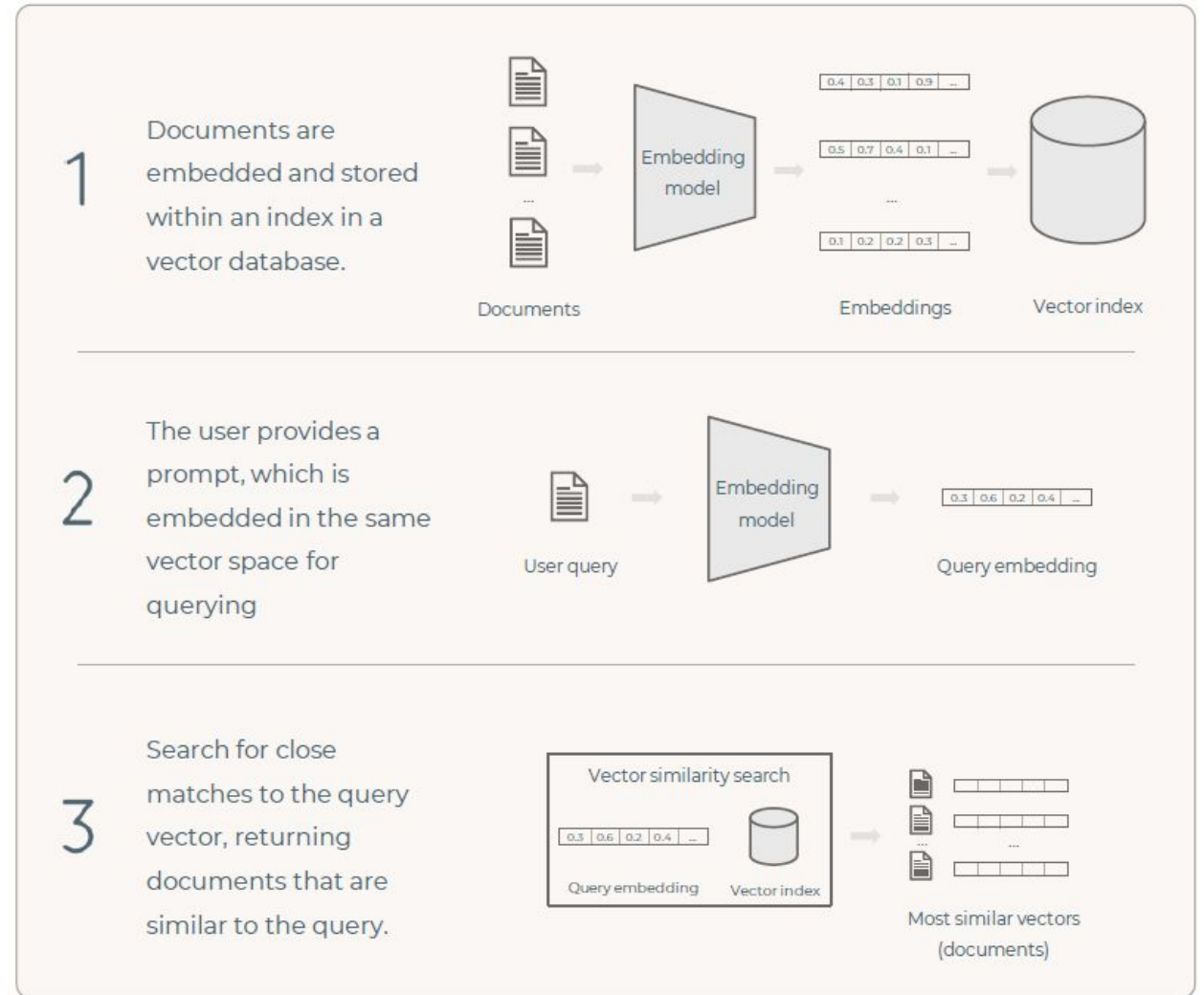
# Encode text descriptions
text_emb = model.encode(
    ["Two dogs in the snow", "A cat on a table", "A picture of London at night"]
)

# Compute similarities
similarity_scores = model.similarity(img_emb, text_emb)
print(similarity_scores)
```


La recuperación de información es el proceso de encontrar datos relevantes dentro de grandes conjuntos de datos.

Su uso principal es ayudar a los usuarios a acceder a la información que necesitan de fuentes digitales.

- **Indexación:** Transformación de datos no estructurados en vectores de alta dimensión (embeddings), que se almacenan. Estos vectores pueden contener tanto información léxica como información semántica.
- **Recuperación:** Una pregunta (query) se convierte en un embedding. Este vector se compara con las incrustaciones almacenadas para encontrar las más similares.



Quantized Retrieval - Binary Search with Scalar (int8) Rescoring

This demo showcases retrieval using [quantized embeddings](#) on a CPU. The corpus consists of 41 million texts from Wikipedia articles.

► Click to learn about the retrieval process

Query for Wikipedia articles

When Abert Einstein was born?

Search Index

☐ Exact Search ☒ Approximate Search

Number of documents to retrieve

Number of documents to retrieve from the binary search

100

Rescore multiplier

Search for 'rescore_multiplier' as many documents to rescore

1

Search

Score ▲	Title ▲	Text ▲
7726	Military history of Jewish Americans	Albert Einstein (1879-1955) was a theoretical physicist and
7611.65	Albert Einstein	Albert Einstein (; 14 March 1879 - 18 April 1955) was a
7356.58	Albert (given name)	Albert Einstein (1879-1955) German-born theoretical physici
7301.69	Alfred Einstein	Alfred Einstein (December 30, 1880February 13, 1952) was a
7240.42	Albert Einstein	Born in the German Empire, Einstein moved to Switzerland :
7185.46	Outline of Albert Einstein	Albert Einstein - deceased German-born theoretical physici
7185.46	Church of the Holy Trinity and Rectory (Middletown, Connecticut)	This house, built in 1902, was the third rectory of the Ho
6970.74	Political views of Albert Einstein	Albert Einstein (1879-1955), a German-born scientist, was
6958.2	Einstein family	Hans Albert Einstein (May 14, 1904 - July 26, 1973) was bo
6939.94	The Infinite Worlds of H. G. Wells	Albert Einstein was born in 1879, so he was in his 20's at
6822.05	1955 in science	April 18 - Albert Einstein (born 1879), German-born theore
6795.2	Hans Albert Einstein	Hans Albert Einstein (May 14, 1904 - July 26, 1973) was a

```
{
  Embed Time: "0.0735 s",
  Quantize Time: "0.0001 s",
  Search Time: "0.1071 s",
  Load Time: "0.1267 s",
  Rescore Time: "0.0005 s",
  Sort Time: "0.0505 s",
  Total Retrieval Time: "0.2849 s"
}
```

Retrieval-Augmented Generation

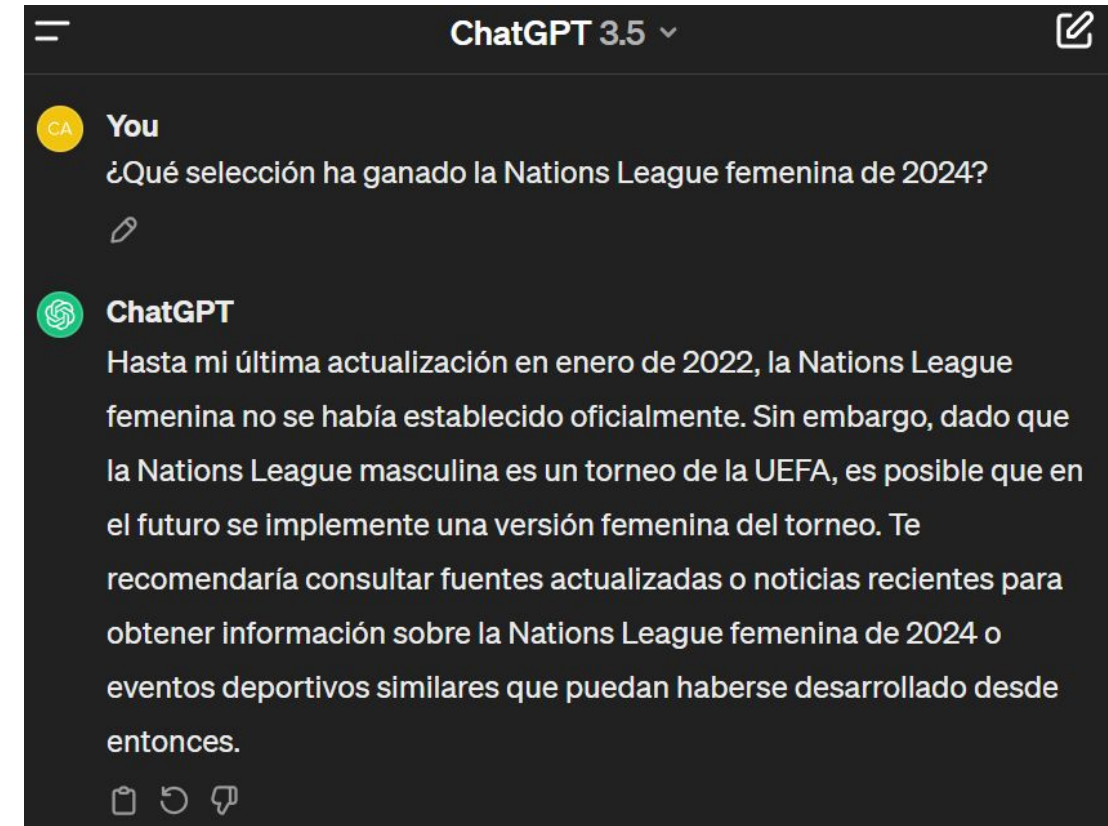
España hace historia y se proclama campeona de la Nations League femenina

La selección femenina de fútbol bate 2-0 a Francia y consigue el segundo título de su historia en seis meses



Las jugadoras de la selección española celebran la conquista de la Liga de Naciones. | Europa Press

Fecha noticia:
28/02/2024



Fecha consulta:
28/02/2024

<https://theobjective.com/deportes/futbol/2024 02 28/espana campeona nations league/>
<https://chat.openai.com/share/bd091e04 e8ae 4a5c 846b e5b88fad69c4>

Retrieval-Augmented Generation (RAG) es una técnica que combina la generación de texto y la recuperación de información para mejorar la precisión y relevancia de las respuestas generadas por modelos de lenguaje.

¿Para qué sirve?

- **Generación de texto mejorada:** Utiliza información recuperada de grandes bases de datos para generar respuestas más precisas y contextualizadas.
- **Respuestas basadas en datos:** Permite que los modelos generen respuestas basadas en datos específicos recuperados en tiempo real, en lugar de depender únicamente de la información aprendida durante el entrenamiento.
- **Manejo de consultas complejas:** Mejora la capacidad de los modelos para manejar consultas complejas o poco comunes al acceder a información adicional relevante.

¿Cómo funciona?

- **Recuperación de información:** Ante una consulta, el sistema primero busca y recupera información relevante de una base de datos o corpus de documentos.
- **Generación de respuesta:** Utiliza la información recuperada para generar una respuesta que combina el conocimiento del modelo con datos específicos y actualizados.

LLM sin contexto

España hace historia y se proclama campeona de la Nations League femenina

La selección femenina de fútbol bate 2-0 a Francia y consigue el segundo título de su historia en seis meses



Las jugadoras de la selección española celebran la conquista de la Liga de Naciones. | Europa Press

ChatGPT 3.5

You

¿Qué selección ha ganado la Nations League femenina de 2024?

ChatGPT

Hasta mi última actualización en enero de 2022, la Nations League femenina no se había establecido oficialmente. Sin embargo, dado que la Nations League masculina es un torneo de la UEFA, es posible que en el futuro se implemente una versión femenina del torneo. Te recomendaría consultar fuentes actualizadas o noticias recientes para obtener información sobre la Nations League femenina de 2024 o eventos deportivos similares que puedan haberse desarrollado desde entonces.

RAG

Microsoft Bing

Búsqueda

Copilot

Bloc De Notas

carlos.rodri...

Protegido

Usted

¿Qué selección ha ganado la Nations League femenina de 2024?

Copilot

Los datos personales y de la empresa están protegidos en este chat

La **selección española femenina de fútbol** se ha proclamado **campeona de la primera edición de la UEFA Nations League** al derrotar por **2-0 a Francia** en la final disputada en el **Estadio de La Cartuja de Sevilla**, sumando su segundo título en seis meses tras haber hecho historia con la conquista del Mundial el pasado verano ¹. ¡Una gran hazaña para España! 🏆 ⚽ es

Más información

1 theobjective.com


2 bing.com


3 estadiodeportivo.com


4 eldesmarque.com


5 larazon.es


1 de 30 respuestas


 Microsoft Bing


 Búsqueda


 Copilot

 Bloc De Notas


carlos.rodri... 


 **Protegido**



 **Usted**


¿Qué selección ha ganado la Nations League femenina de 2024?


 **Copilot**


 Los datos personales y de la empresa están protegidos en este chat


La **selección española femenina de fútbol** se ha proclamado **campeona de la primera edición de la UEFA Nations League** al derrotar por **2-0 a Francia** en la final disputada en el **Estadio de La Cartuja de Sevilla**, sumando su segundo título en seis meses tras haber hecho historia con la conquista del Mundial el pasado verano ¹. ¡Una gran hazaña para España! 🏆 🌐 ES


Más información





1  [theobjective.com](#)

2  [bing.com](#)

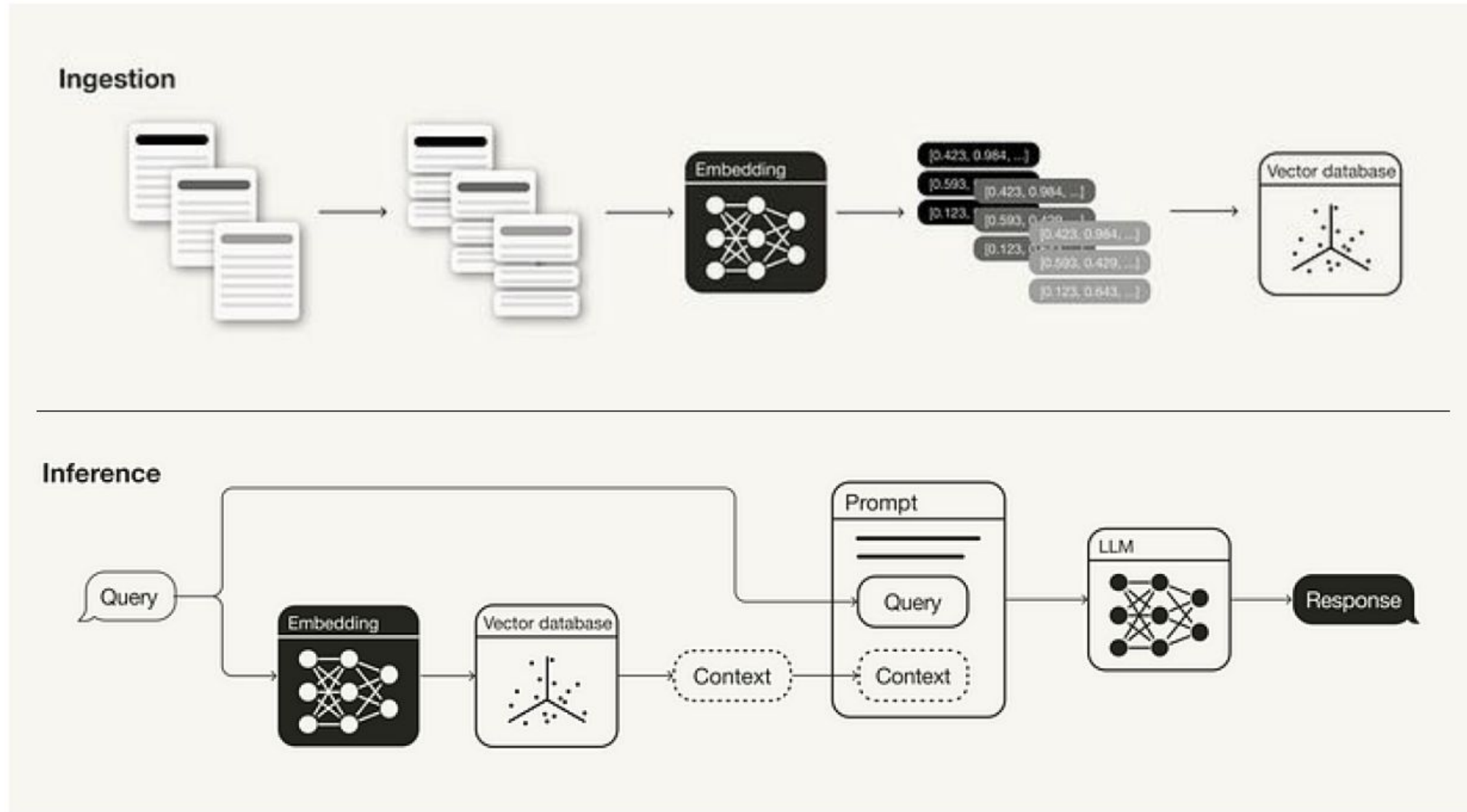
3  [estadiodeportivo.com](#)

4  [eldesmarque.com](#)

5  [larazon.es](#)

1 de 30 respuestas



Casos de éxito



Forbes



Ranking Educativo
Innovatec



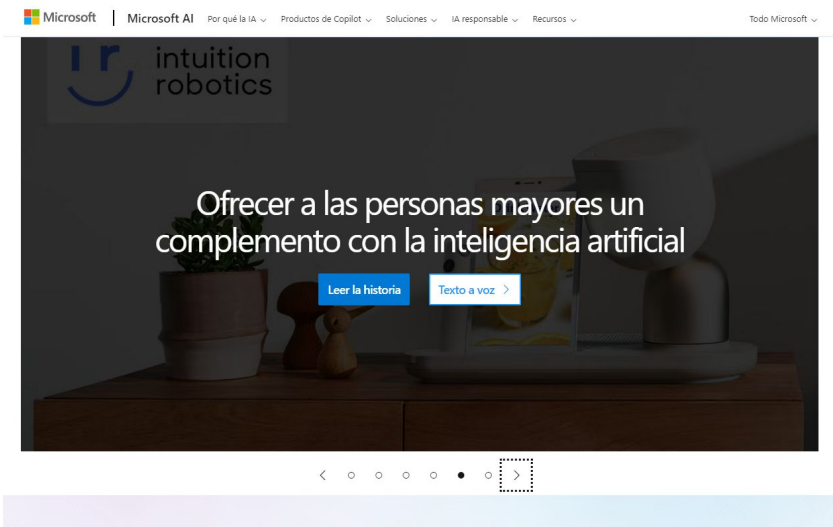
EL MUNDO



Los **success stories** (historias de éxito) son estudios de caso o relatos detallados que describen cómo una empresa o una organización ha implementado una solución tecnológica para alcanzar resultados significativos. Estas historias generalmente incluyen:

- **Desafíos:** Una descripción del problema o desafío que enfrentaba la organización.
- **Solución:** Una explicación de la solución implementada, incluyendo tecnologías y servicios específicos.
- **Resultados:** Datos y métricas que demuestran el impacto positivo de la solución.

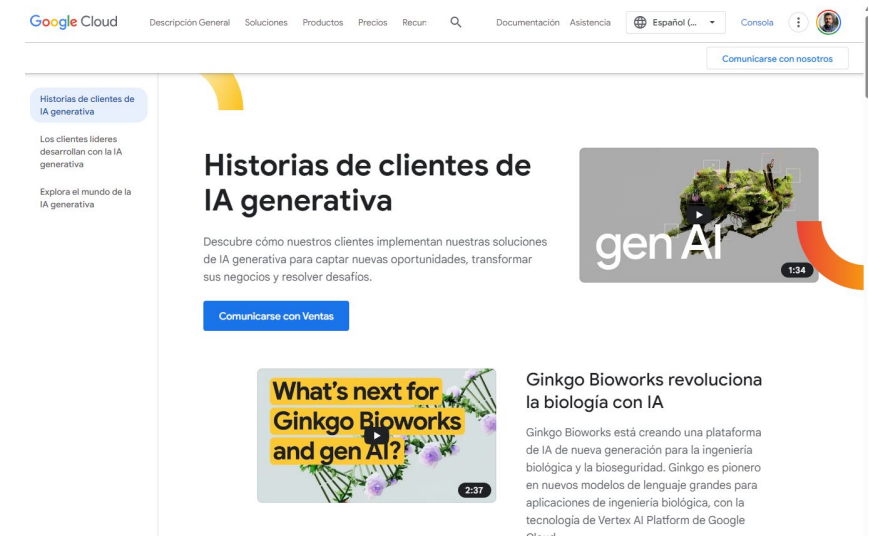
Microsoft



AWS



Google



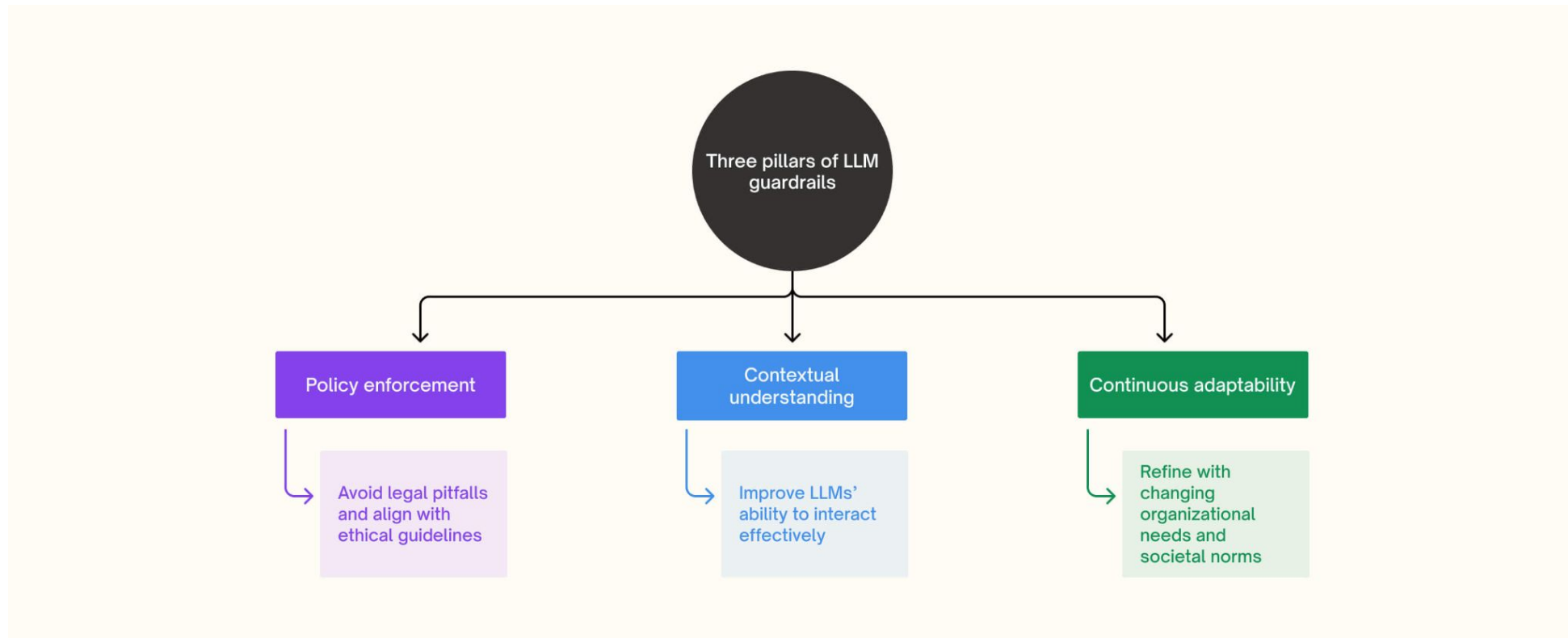
Links:

- <https://www.microsoft.com/es-es/ai/ai-customer-stories>
- https://aws.amazon.com/es/ai/generative-ai/customers/?customer-references-cards.sort-by=item.additionalFields.sortDate&customer-references-cards.sort-order=desc&awsf.customer-references-location=*all&awsf.customer-references-industry=*all
- <https://cloud.google.com/ai/generative-ai/stories?hl=es-419>

Desafíos futuros

- **Sesgos y Equidad**
 - **Persistencia de sesgos:** A pesar de los esfuerzos, los LLM pueden perpetuar sesgos inherentes en los datos de entrenamiento.
 - **Equidad en representación:** Asegurar que todas las comunidades y grupos estén representados de manera justa en los datos.
- **Privacidad y Seguridad**
 - **Protección de datos:** Garantizar que los datos utilizados sean manejados de manera que protejan la privacidad de los usuarios.
 - **Ciberseguridad:** Evitar que los LLM sean explotados para crear contenido malicioso como phishing, deepfakes y desinformación.
- **Sostenibilidad**
 - **Consumo energético:** Los LLM requieren vastas cantidades de recursos computacionales, planteando preocupaciones sobre el impacto ambiental.
 - **Eficiencia de recursos:** Desarrollar modelos más eficientes que operen con menor consumo energético sin sacrificar rendimiento.
- **Interpretabilidad y Transparencia**
 - **Caja negra:** La complejidad de los LLM dificulta la comprensión de cómo llegan a ciertas conclusiones.
 - **Transparencia en decisiones:** Mejorar la interpretabilidad de los modelos para que los humanos puedan entender y confiar en sus decisiones.
- **Regulación y Ética**
 - **Marcos regulatorios:** Desarrollo de regulaciones y políticas que guíen el uso ético y responsable de la IA generativa.
 - **Responsabilidad ética:** Asegurar que las aplicaciones de IA respeten los derechos y la dignidad humana.
- **Escalabilidad y Adaptabilidad**
 - **Eficiencia en el escalado:** Desarrollar técnicas que permitan escalar modelos sin incurrir en costos prohibitivos.
 - **Adaptación continua:** Los modelos deben adaptarse continuamente a nuevos datos y contextos sin necesidad de un reentrenamiento completo.

En el contexto de los LLM, los *guardrails* son un conjunto de reglas, limitaciones, y protocolos que sirven para controlar el comportamiento y la salida de los LLM.



Link: [LLMs Guardrails Guide: What, Why & How | Attri AI Blog | Attri.ai Blog](#)

- **Éticos:** Limitaciones para prevenir resultados discriminatorios, sesgados o dañinos, asegurando que los LLMs operen dentro de normas sociales y morales aceptadas.
- **Cumplimiento:** Alineación de los resultados del modelo con estándares legales, crucial en salud, finanzas y servicios legales, incluyendo la protección de datos y privacidad.
- **Contextuales:** Refinan la comprensión del modelo sobre lo relevante y aceptable en contextos específicos, evitando textos inapropiados.
- **Seguridad:** Protección contra amenazas de seguridad internas y externas, evitando manipulación del modelo para divulgar información sensible o propagar desinformación.
- **Adaptativos:** Evolución continua de los rieles junto con el modelo, asegurando alineación con estándares éticos y legales.



Comparando LLMs: LMSYS Chatbot Arena Leaderboard

[Arena \(battle\)](#)

[Arena \(side-by-side\)](#)

[Direct Chat](#)

[Leaderboard](#)

[About Us](#)

LMSYS Chatbot Arena Leaderboard

[Vote](#) | [Blog](#) | [GitHub](#) | [Paper](#) | [Dataset](#) | [Twitter](#) | [Discord](#)

LMSYS [Chatbot Arena](#) is a crowdsourced open platform for LLM evals. We've collected over 1,000,000 human pairwise comparisons to rank LLMs with the [Bradley-Terry model](#) and display the model ratings in Elo-scale. You can find more details in our [paper](#).

Arena

Full Leaderboard

Total #models: 105. Total #votes: 1,205,934. Last updated: 2024-06-02.

NEW! View leaderboard for different categories (e.g., coding, long user query)! This is still in preview and subject to change.

Code to recreate leaderboard tables and plots in this [notebook](#). You can contribute your vote at [chat.lmsys.org](#)!

*Rank (UB): model's ranking (upper-bound), defined by one + the number of models that are statistically better than the target model. Model A is statistically better than model B when A's lower-bound score is greater than B's upper-bound score (in 95% confidence interval). See Figure 3 below for visualization of the confidence intervals of model scores.

Category		Overall Questions					
Overall		#models: 105 (100%)		#votes: 1,205,934 (100%)			
Rank* (UB)	Model	Arena Elo	95% CI	Votes	Organization	License	Knowledge Cutoff
1	GPT-4o-2024-05-13	1287	+4/-4	32181	OpenAI	Proprietary	2023/10
2	Gemini-1.5-Pro-API-0514	1267	+5/-4	25519	Google	Proprietary	2023/11
2	Gemini-Advanced-0514	1266	+5/-5	27225	Google	Proprietary	Online
4	Gemini-1.5-Pro-API-0409-Preview	1257	+3/-3	55731	Google	Proprietary	2023/11
4	GPT-4-Turbo-2024-04-09	1256	+2/-2	59891	OpenAI	Proprietary	2023/12
5	GPT-4-1106-Preview	1251	+2/-3	80067	OpenAI	Proprietary	2023/4
6	Claude-3-Opus	1248	+2/-2	123645	Anthropic	Proprietary	2023/8

Figure 1: Fraction of Model A Wins for All Non-tied A vs. B Battles

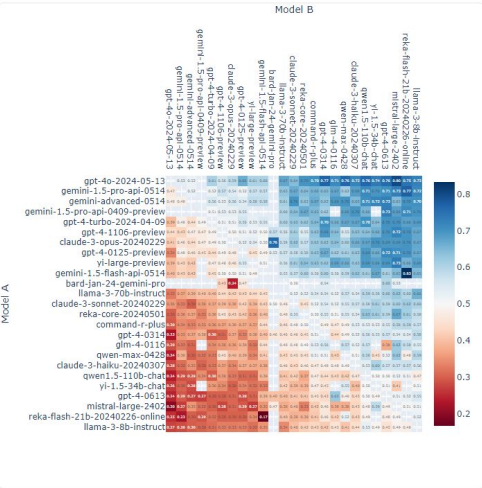


Figure 3: Confidence Intervals on Model Strength (via Bootstrapping)

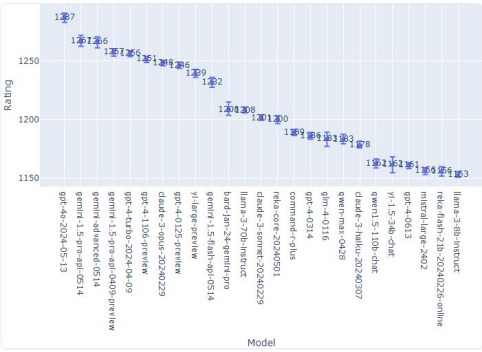


Figure 2: Battle Count for Each Combination of Models (without Ties)

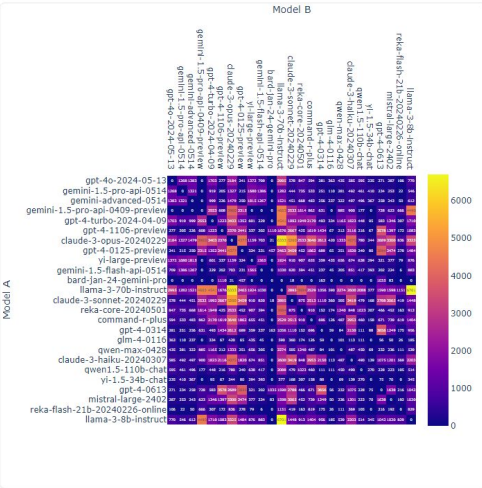
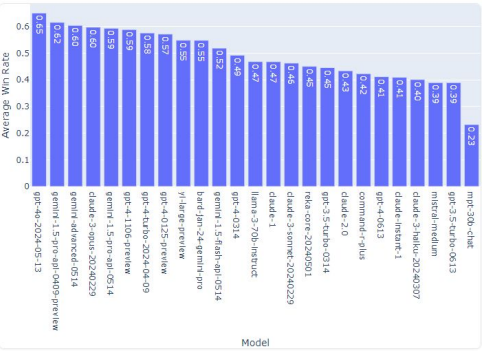


Figure 4: Average Win Rate Against All Other Models (Assuming Uniform Sampling and No Ties)



<https://chat.lmsys.org/?leaderboard>

The screenshot displays a Google Colab notebook interface. The top bar shows the notebook title 'HF_bias_evaluation.ipynb' and various menu options like 'Archivo', 'Editar', 'Ver', 'Insertar', 'Entorno de ejecución', 'Herramientas', and 'Ayuda'. The left sidebar contains a table of contents with sections such as 'Evaluating Bias and Toxicity in Language Models', 'Choosing a model', 'Evaluating Toxicity', 'Loading Real Toxicity Prompts', 'Prompting the Model', 'Evaluating toxicity of model outputs', 'Evaluating Regard', 'Loading BOLD prompts', 'Prompting our Model', 'Calculating Regard', 'HONEST', 'Loading HONEST prompts', 'Prompting the Model', and 'Calculating the HONEST score'. The main content area is titled 'Evaluating Bias and Toxicity in Language Models' and contains the following text:

In this notebook, we'll see how to evaluate different aspects of bias and toxicity of large language models hosted on [Hugging Face Transformers](#). We will cover three types of bias evaluation, which are:

- **Toxicity:** aims to quantify the toxicity of the input texts using a pretrained hate speech classification model.
- **Regard:** returns the estimated language polarity towards and social perceptions of a demographic (e.g. gender, race, sexual orientation).
- **HONEST score:** measures hurtful sentence completions based on multilingual hate lexicons.

The workflow of the evaluations described above is the following:

- Choosing a language model for evaluation (either from the [Hugging Face Hub](#) or by training your own)
- Prompting the model with a set of predefined prompts
- Running the resulting generations through the relevant metric or measurement to evaluate its bias.

First things first: you need to install [Hugging Face Transformers](#), [Datasets](#) and [Evaluate](#)!

If you're opening this notebook locally, make sure your environment has an install from the last version of those libraries.

```
[ ] !pip install datasets transformers evaluate -q
```

Next, the 'Choosing a model' section begins, explaining that the steps depend on being able to *prompt* your model to evaluate its *generations*. It mentions consulting the [Hugging Face Hub](#) for models capable of this and provides a prompt for [GPT-2](#).

The 'Evaluating Toxicity' section follows, stating that evaluating toxicity can help understand a model's behavior and that it can be done by prompting the model with a set of prompts and then evaluating the toxicity of its generations.

The 'Loading Real Toxicity Prompts' section explains that the notebook will use [Real Toxicity Prompts](#), a dataset of prompts created specifically for this purpose, containing 100,000 (!) prompts stratified across 4 levels of toxicity.

Link: <https://colab.research.google.com/drive/1-HDJUcPMKEF-E7Hapih0OmA1xTW2hdAv#scrollTo=yX8ciyVWKiuO>

Próximos pasos



Forbes



Ranking Educativo
Innovatec



EL MUNDO



- Repaso y lectura de los conceptos tratados en la 4ª sesión.
- Actividad evaluable:
 - Test multirespuesta tema 4 (individual).
 - Caso práctico (equipos TFM).
- Dudas y preguntas, vía *tablero de discusión* o email.

OBS Business
School



Planeta Formación y Universidades