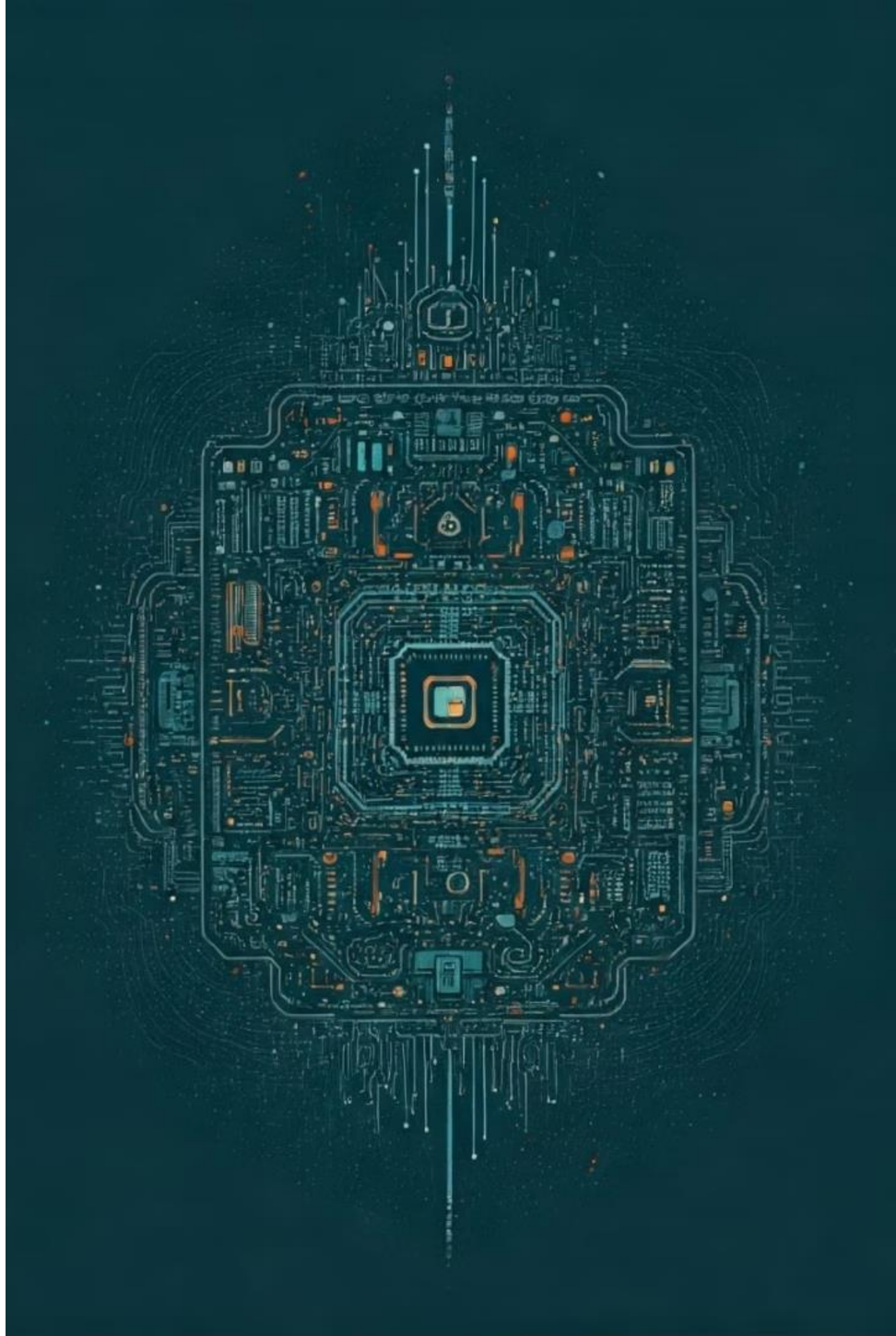
Modelos del lenguaje

by Ivan Ruiz Rube



Contenidos

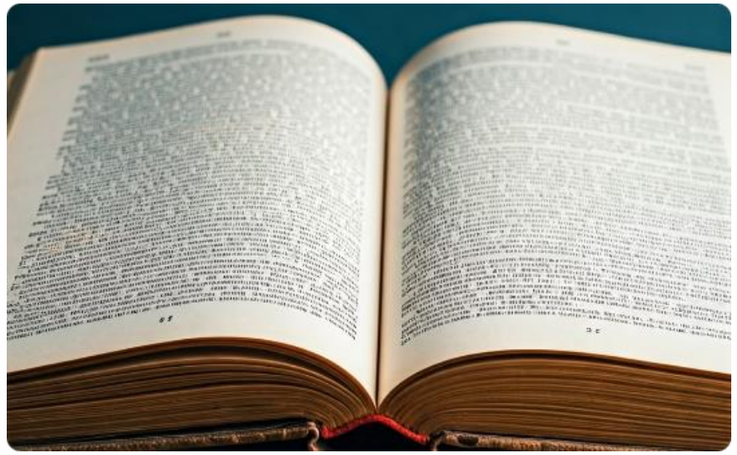
- Anatomía de los modelos de lenguaje
- LLM as a Black Box
- Tipos de modelos de lenguaje
- La batalla de los modelos
- Ventajas e inconvenientes

Anatomía de los modelos de lenguaje

Modelos de lenguaje

Machine Learning

Son **modelos** de IA entrenados con grandes conjuntos de datos de texto, aprendiendo las reglas, patrones y estructuras del lenguaje.



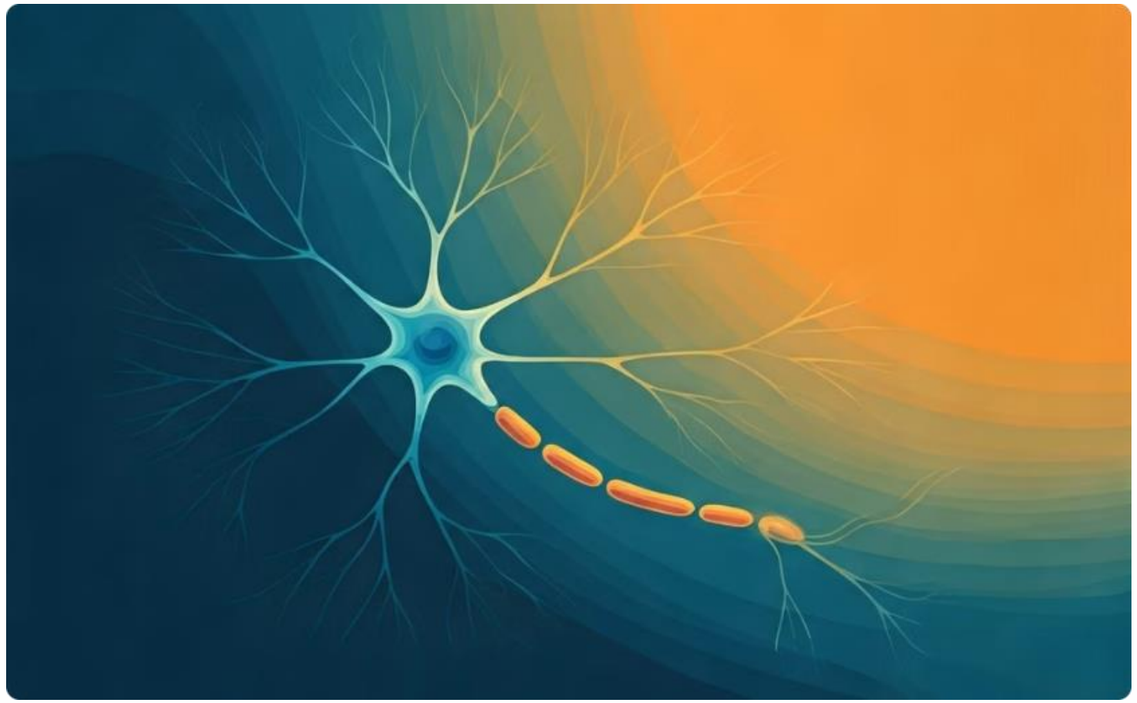
Procesamiento del Lenguaje Natural

Diseñados para entender, generar y manipular texto en lenguaje natural.

Redes neuronales artificiales

Algoritmos inspirados en el funcionamiento del cerebro humano, diseñados para reconocer patrones y procesar datos de manera eficiente.

Redes neuronales



Estructura

Las redes neuronales se basan en la estructura de las neuronas biológicas. Estas se componen de dendritas, un núcleo y un axón. axón.

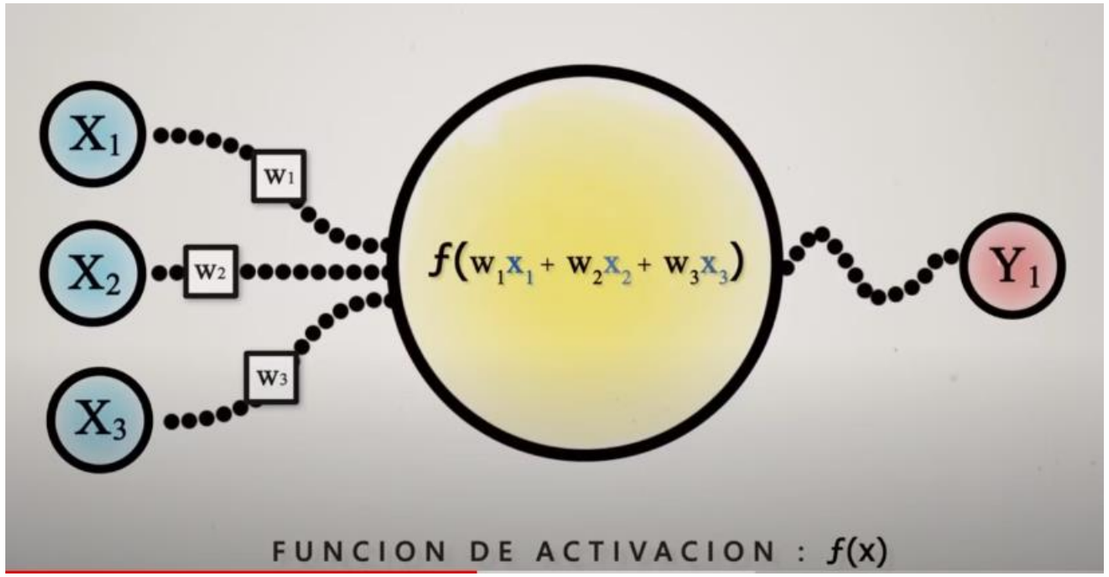
Interconexión

Las neuronas son células que se conectan entre sí para formar una red. una red. Esta red permite la recepción, procesamiento y transmisión de transmisión de información a través de señales químicas y eléctricas. eléctricas.

Neurona artificial

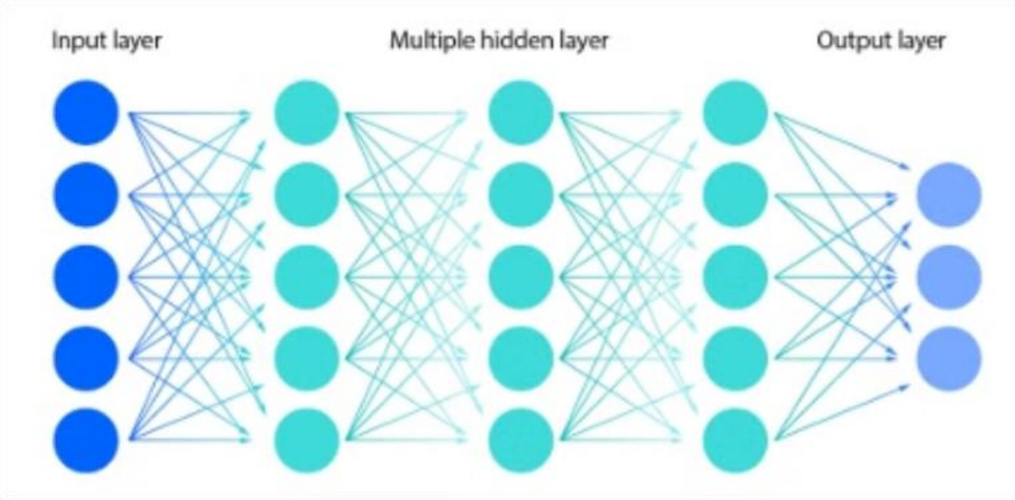
Es un nodo de procesamiento que recibe entradas, realiza una una operación matemática sobre ellas y produce una salida.

La operación matemática, llamada función de activación, determina el comportamiento de la neurona.



https://www.youtube.com/watch?v=MRIv2IwFTPg

Deep Learning



Capa de entrada

Recibe los datos de entrada, como texto texto o imágenes.

Capas intermedias

Procesan la información y extraen características relevantes en múltiples múltiples niveles.

Capa de salida

Genera la salida final, como una predicción o una clasificación.

https://www.ibm.com/es-es/topics/neural-networks

Arquitecturas de redes neuronales

Multilayer Perceptron (MLP)

Las MLP son redes neuronales que utilizan capas ocultas para aprender funciones complejas. Son útiles para el reconocimiento de patrones, como imágenes o texto.

Convolutional Neural Network (CNN)

Las CNN son redes neuronales especializadas para el análisis de imágenes. Emplean operaciones de convolución para extraer características relevantes, como bordes y texturas.

Recurrent Neural Network (RNN)

Las RNN procesan secuencias de datos, como el lenguaje natural o el reconocimiento de voz. Permiten "recordar" información previa en la secuencia.

Generative Adversarial Network (GAN)

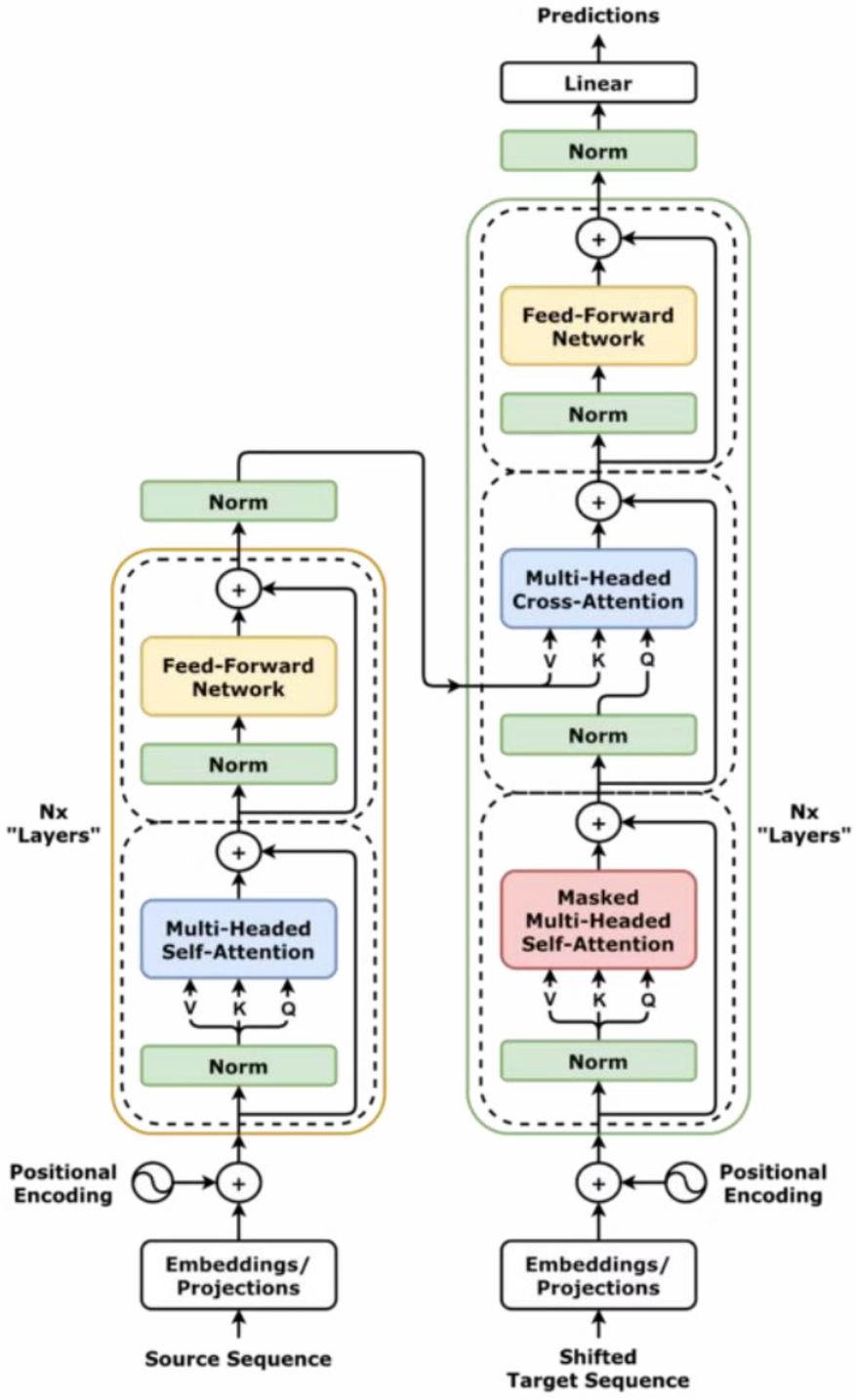
Las GAN consisten en dos redes neuronales que compiten para compiten para generar datos realistas. Un generador crea datos y crea datos y un discriminador los distingue.

Arquitectura de los modelos de lenguaje

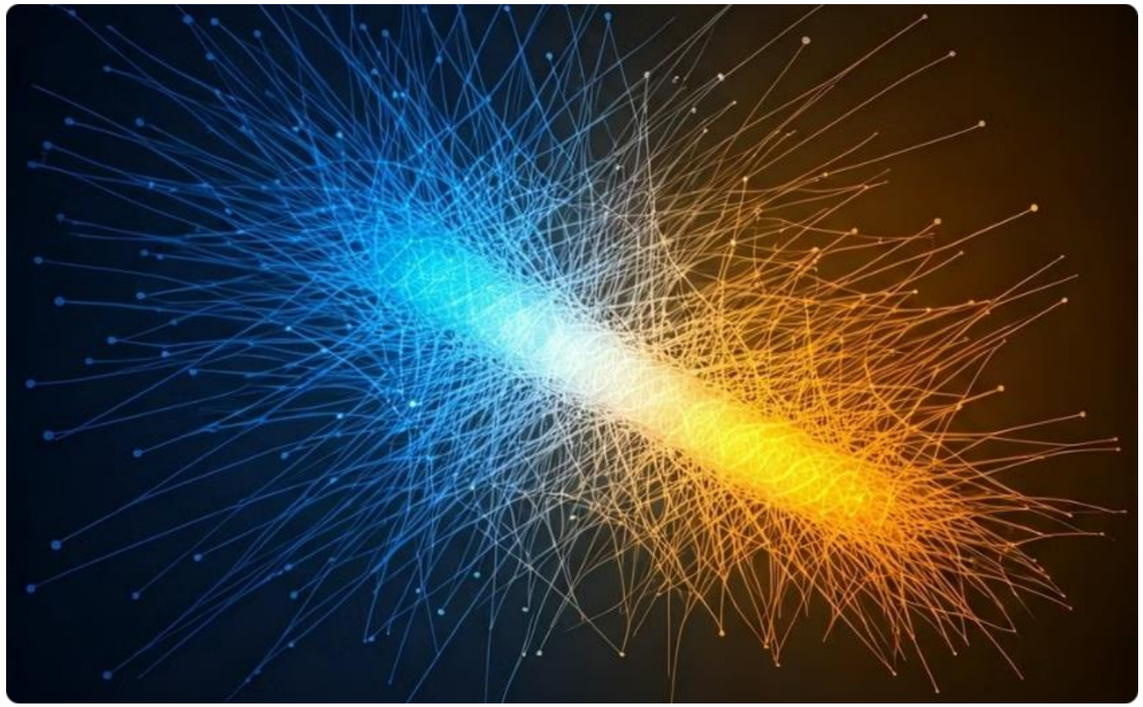
Transformers

Los modelos de lenguaje aprovechan la arquitectura *Transformer*, la cual permite permite capturar relaciones contextuales de forma eficiente y mejorar considerablemente la comprensión del lenguaje natural.

GPT = Generative Pretrained Transformer



Procesos clave



Entrenamiento

Los modelos de lenguaje se entrenan mediante redes neuronales que ajustan que ajustan los pesos de las conexiones entre neuronas. Para este entrenamiento se requiere una gran cantidad de datos.

Inferencia

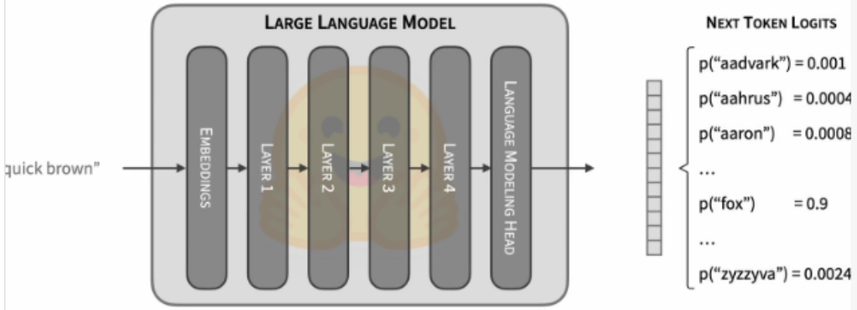
Durante la inferencia, los modelos de lenguaje utilizan la generación auto-generación auto-regresiva para predecir la siguiente palabra en una secuencia. El proceso termina cuando se alcanza una longitud máxima o se máxima o se genera un token especial.

Tokens

Un token es la unidad básica de texto que los modelos de lenguaje utilizan para procesar y generar lenguaje.

Número de Tokens≈Número de Palabras×1.3

Funcionamiento básico



huggingface

Generation with LLMs

We're on a journey to advance and democratize artificial intelligence through open source and open science.

Embeddings

Los embeddings son **representaciones matemáticas** del significado de las palabras o frases.

Cada palabra o frase se convierte en un **vector numérico**, un conjunto de números de números que refleja su significado semántico.

Estas representaciones permiten realizar **operaciones matemáticas** para comparar el significado de palabras o frases.

Embeddings

Ejemplo

Amarillo -> (255, 255, 0)

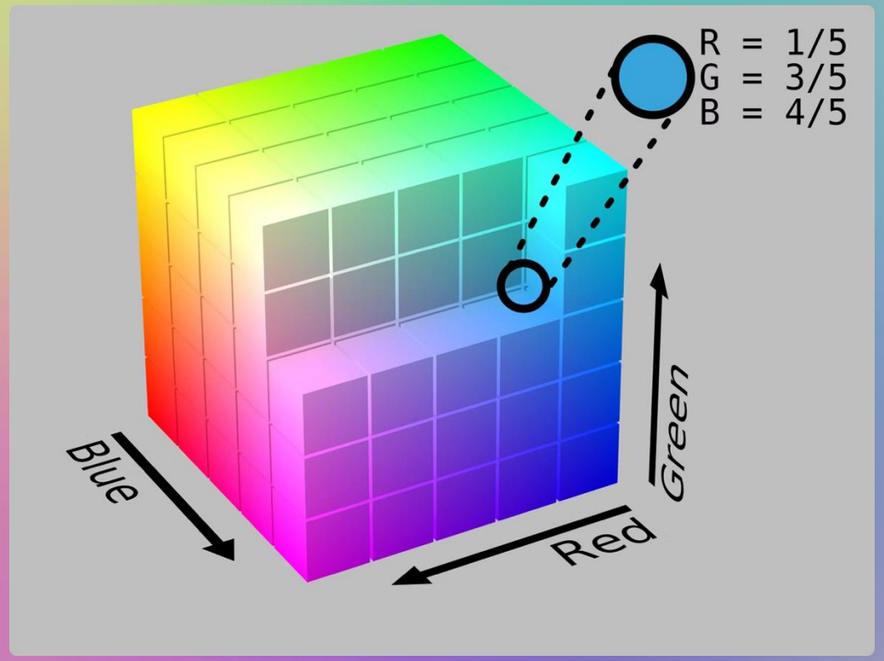
Verde -> (0, 255, 0)

Azul -> (0, 0, 255)

Operaciones matemáticas

Amarillo – Verde + Azul = Magenta

(255, 255, 0) - (0, 255, 0) + (0, 0, 255) = (255, 0, 255)



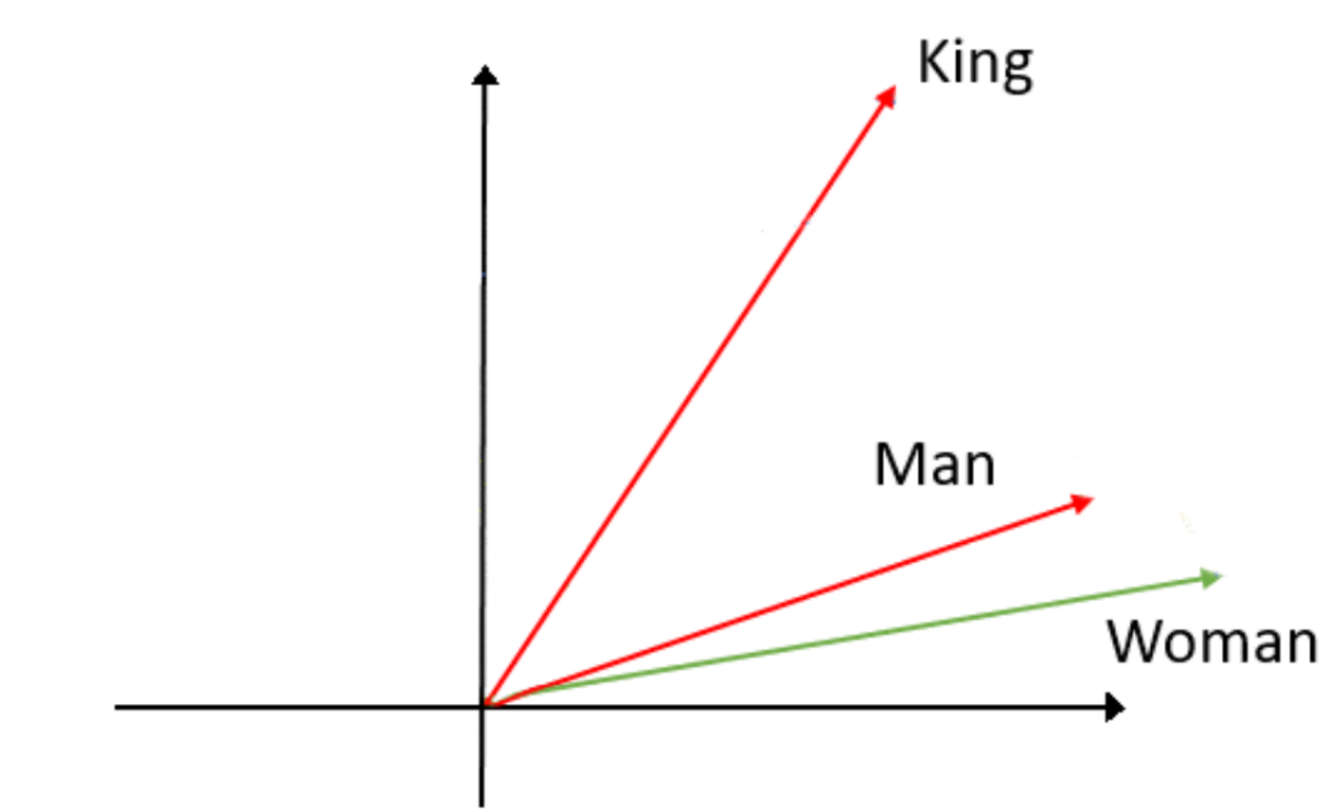
Visualización de embeddings

projector.tensorflow.org

Embedding projector - visualization of high-dimensional data

Visualize high dimensional data.

King - Man + Woman = Queen



LLM as a Black Box

LLM as a black box

1

Text input

El usuario introduce el texto que desea que el LLM procese. Los modelos de lenguaje procesan el texto como una secuencia de **tokens**.

2

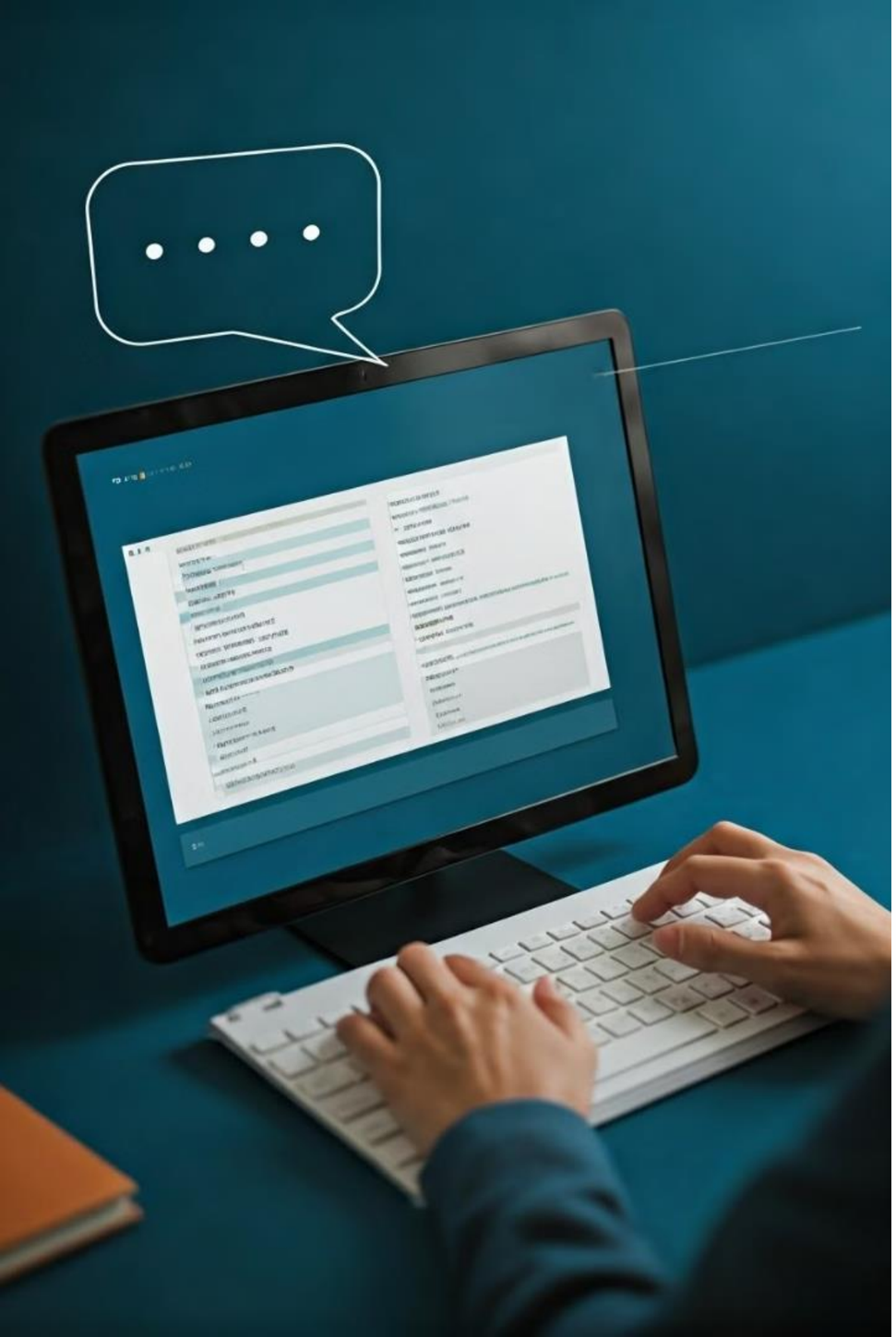
LLM inference engine

El motor de inferencia utiliza el modelo del lenguaje para procesar el texto de entrada, convirtiendo los tokens en embeddings y calculando las probabilidades de salida para cada token.

3

Text output

El modelo del lenguaje genera texto como salida, a partir de la predicción predicción de tokens, creando un texto coherente con sentido.



Text input/output

Entrada

Un **prompt** es una instrucción o pregunta que se le da al modelo de de lenguaje.

Salida

El modelo genera una secuencia de de palabras y frases coherentes con la con la entrada.

Calidad

La calidad de la respuesta generada generada depende del diseño del prompt.

Precisión

Es importante diseñar los prompts con con precisión y consistencia.

Motor de inferencia

Ejecución de modelos

Ejecutan modelos de de lenguaje preentrenados.

Entrada y salida

Procesan entradas nuevas para generar generar salidas relevantes.

Escalabilidad

Gestionan múltiples múltiples solicitudes solicitudes simultáneamente.

API REST

Facilitan la integración dentro dentro de aplicaciones.

Formato de los modelos

Los modelos de lenguaje generalmente se almacenan en forma de archivos binarios que contienen los **pesos** del modelo

PyTorch (.pt o .pth)

PyTorch es conocido por su flexibilidad y facilidad de uso.

TensorFlow (ckpt y SavedModel)

TensorFlow es una plataforma de aprendizaje automático de código abierto.

Keras (.h5)

Keras, la API de alto nivel de TensorFlow, facilita el desarrollo rápido de prototipos.

Open Neural Network Exchange (.onnx)

ONNX es un formato de modelo de aprendizaje automático abierto que permite la interoperabilidad entre diferentes frameworks.

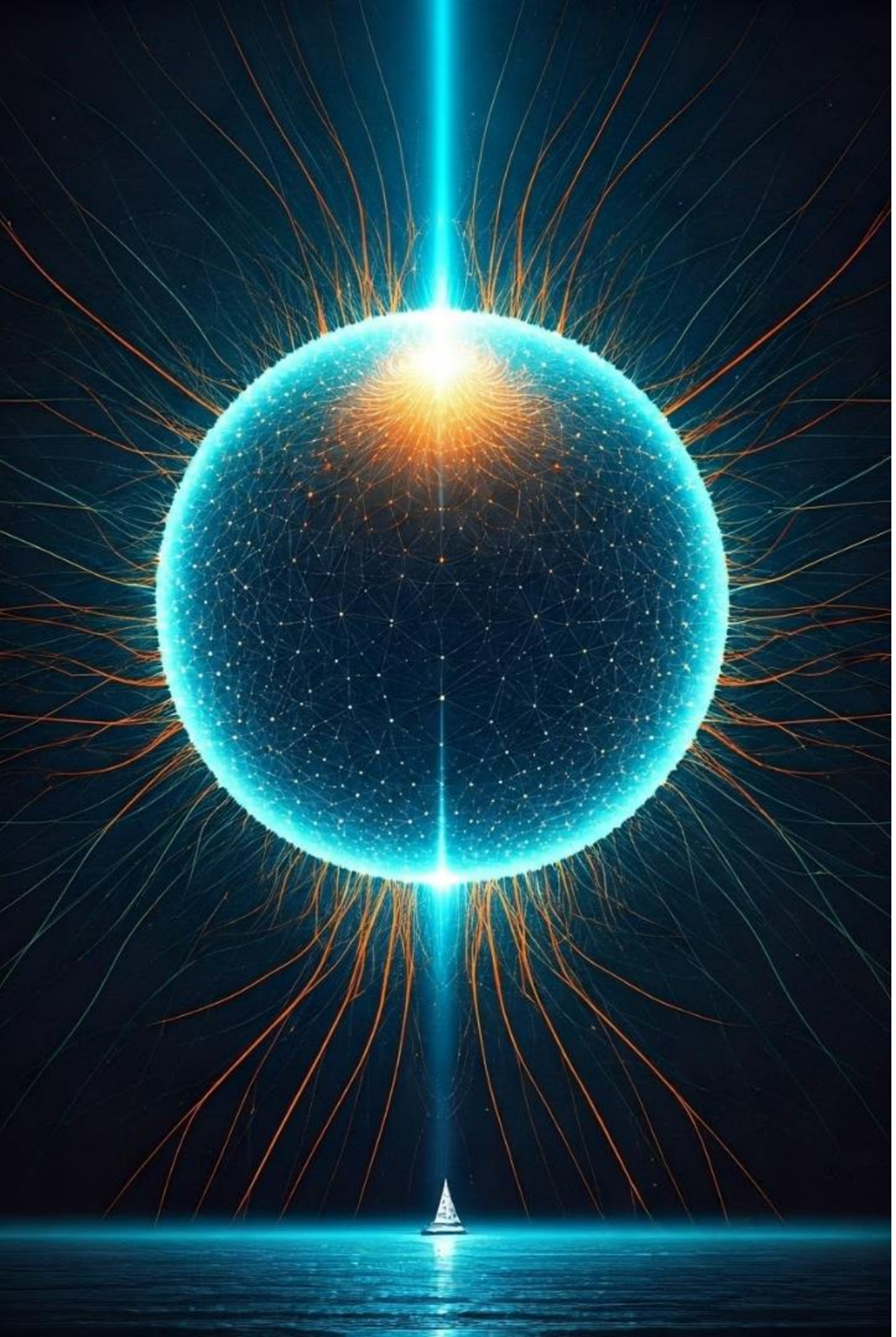
GGML/GGUF

Optimizan el almacenamiento y la inferencia en dispositivos con recursos limitados.

Hugging Face Transformers (.bin)

Hugging Face Transformers proporciona una biblioteca de modelos preentrenados.

Tipos de modelos de lenguaje



Foundational models

Capacidad Generalizada

Entrenados en vastos corpus de datos sin etiquetar. Realizan una amplia gama de tareas generales.

Inversión Significativa

Desarrollo desde cero requiere millones de dólares. Infraestructura y recursos computacionales extensivos.

Ejemplos Notables

GPT-4 y Llama3-pretrained destacan como modelos fundacionales ampliamente reconocidos.

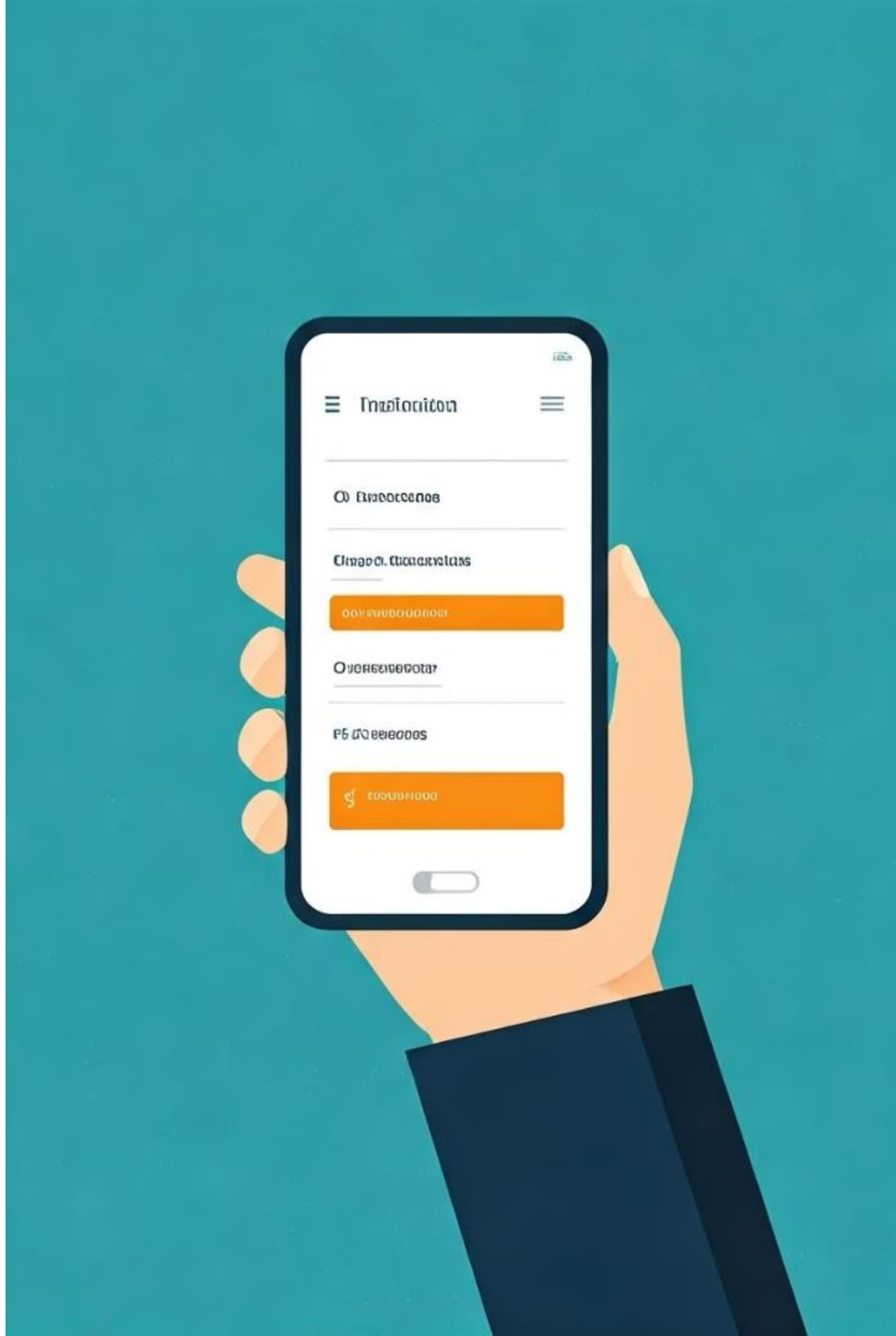
Instruct models

Modelos que han sido entrenados con conjuntos de datos que incluyen instrucciones específicas, lo que les permite seguir indicaciones de manera más manera más precisa y coherente.

Estos modelos son especialmente útiles para aplicaciones interactivas donde el donde el modelo necesita comprender y ejecutar instrucciones detalladas del detalladas del usuario.

Mejor capacidad para seguir órdenes y realizar tareas específicas.

Respuestas más alineadas con las expectativas del usuario.



Conversational (Chat) Models

Estos modelos están instruidos para interactuar con usuarios para mantener mantener conversaciones fluidas y coherentes.

Se usan en asistentes virtuales, chatbots y sistemas de soporte al cliente.

Son capaces de mantener el contexto, ofreciendo respuestas naturales y lógicas, lógicas, adaptándose al estilo y preferencias del usuario.

Domain-Specific Models

Especialización

Modelos enfocados en campos específicos. Ofrecen mayor precisión en áreas concretas.

Ventajas

Comprensión profunda de terminología terminología y contextos especializados. especializados. Mejoran la toma de decisiones en sectores específicos.

Ejemplos

LegalGPT para derecho, ClinicalCamel en medicina, y FinBERT en finanzas.

Modelos multimodales

Los modelos de lenguaje de última generación no solo procesan texto, sino que sino que pueden interactuar con diferentes formatos de datos.

Estos modelos pueden recibir imágenes, audios, vídeos y texto como entrada, y entrada, y generar respuestas en cualquier combinación de estos formatos. formatos.

Los LLM multimodales representan un avance significativo en la inteligencia artificial, abriendo nuevas posibilidades para la interacción hombre-máquina.

Ejemplos: GPT-4o, Gemini 1.5 Pro

La batalla de los modelos

Hitos más destacados

Arquitectura de Transformers

1

La arquitectura Transformer revolucionó el procesamiento del lenguaje natural en 2017, permitiendo a los modelos procesar información en paralelo y manejar dependencias a largo plazo de manera más eficiente.

Modelo BERT

2

BERT (Bidirectional Encoder Representations from Transformers) de 2018 demostró la importancia del preentrenamiento bidireccional, mejorando el rendimiento en tareas de NLP.

GPT-3

3

GPT-3 marcó un hito en 2020 con su tamaño y capacidad, demostrando demostrando que un modelo grande y generalista podía generar texto texto coherente y de alta calidad en múltiples idiomas y tareas.

ChatGPT: Fine tuning para Instrucciones

4

El avance en el ajuste fino de modelos para instrucciones en 2022 mejoró mejoró la capacidad de los LLM para seguir instrucciones de manera manera precisa y realizar tareas más especializadas.

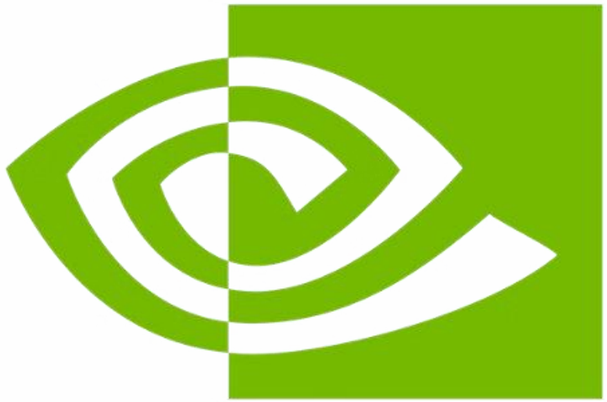
Modelos Multimodales

5

En 2024, GPT-4o y otros modelos multimodales permitieron la integración integración de múltiples tipos de información.

Big Players

OpenAI



NVIDIA®

OpenAI

Pioneros en modelos GPT y la herramienta ChatGPT, liderando la liderando la innovación en IA conversacional.

Nvidia

Líder en el diseño y desarrollo de unidades de procesamiento gráfico (GPU) avanzadas que han impulsando esta revolución de la GenAI

Otros Actores Importantes

Microsoft

Colaboración estrecha con OpenAI. Integración de IA en productos como Office y Azure.

Anthropic

Desarrolladores del modelo Claude. Claude. Enfoque en IA segura y ética. ética.

Alibaba

Gigante chino de comercio electrónico electrónico que invierte en IA y desarrollo de LLM.

Meta

Desarrollo de LLM "open source", incluyendo el modelo LLaMA .

Mistral

Empresa francesa especializada en el desarrollo de modelos de código abierto. Referente en la UE.

Google

Líder en la innovación en IA con búsqueda y productos

Clasificación de Modelos

Parámetros

Varían desde millones hasta billones. Influyen en la capacidad capacidad y complejidad del modelo.

Ventana de Contexto

Desde cientos hasta decenas de de miles de tokens. Afecta la comprensión de textos largos. largos.

Especialización

Desde modelos generalistas hasta altamente especializados especializados en dominios específicos.

Licencia de uso

Desde modelos privativos hasta hasta modelos (de código abierto)

Parámetros

1

SLM (Small Language Model)

Aproximadamente 1 billón de parámetros. Ideal para aplicaciones con recursos limitados.

2

MLM (Medium Language Model)

Alrededor de 8 billones de parámetros. Equilibrio entre rendimiento y eficiencia.

3

LLM (Large Language Model)

Cerca de 70 billones de parámetros. Capacidad superior para para tareas complejas.

Ventana de Contexto

Cantidad de información de entrada y salida para una inferencia del modelo.

1. **Ventana corta**: Procesa entre **512 y 2048 tokens (1 a 5 5 páginas** de texto en español). Ideal para respuestas breves breves o consultas simples.
2. **Ventana media**: Maneja entre **2048 y 8192 tokens (5 a 20 a 20 páginas)**. Adecuada para diálogos multi-turno y textos textos de longitud moderada.
3. **Ventana larga**: Procesa **más de 8K tokens (más de 20 páginas)**. Perfecta para manejar documentos extensos o análisis de información compleja.

Especialización

Los LLM generalistas son capaces de realizar una amplia gama de tareas, mientras que los especializados son diseñados para un propósito específico.

Hard Prompts

Evalúa la capacidad del modelo para manejar entradas complejas o ambiguas. ambiguas.

Instruction Following

Mide la capacidad del modelo para seguir y ejecutar instrucciones precisas.

Coding

Evalúa la habilidad del modelo para generar y manipular código.

Math

Mide la capacidad del modelo para resolver problemas matemáticos.

Multi-Turn Conversations

Evalúa la capacidad del modelo para mantener conversaciones coherentes y contextualizadas.

Long Queries

Mide la capacidad del modelo para comprender y procesar entradas largas.

Razonamiento Lógico

Evalúa la capacidad del modelo para razonar de forma lógica y sacar conclusiones válidas.

Memoria y Contexto

Mide la capacidad del modelo para recordar información previa y mantener el contexto durante una conversación.

Modelos Privativos vs. Abiertos

Disponibilidad:

Privativos: Desarrollados por empresas, con acceso restringido bajo API
API (habitualmente de pago) y código cerrado.

Abiertos: Disponible para descarga e instalación (on-premises o en la nube).

Acceso y Licenciamiento:

Privativos: Requieren pago o subscripción, con licencias restrictivas.

Abiertos: Gratuitos y accesibles bajo licencias abiertas, permiten personalización.

Control y Personalización:

Privativos: Limitada personalización, controlado por la empresa.

Abiertos: Control total del usuario para modificar y adaptar.

Calidad y costes:

Privativos: Alta calidad, pero costos elevados y restricciones.

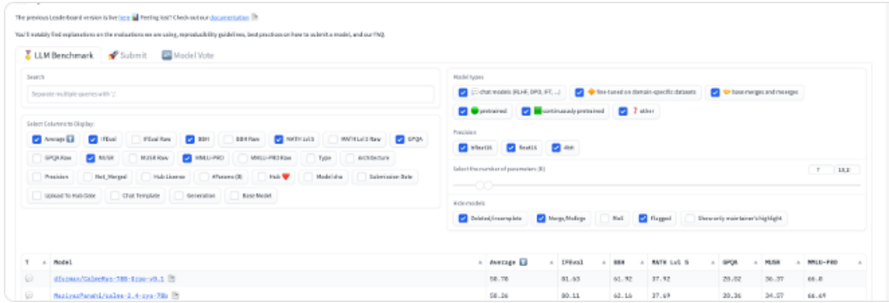
Abiertos: *Menos potentes y más económicos de implantar.*

Transparencia:

Privativos: Poca o ninguna información sobre los datos de entrenamiento

Abiertos: *Ofrecen información suficientemente detallada sobre los datos usados datos usados para entrenar el sistema?*, o bien, *Los datos con los que ha sido ha sido entrenado el modelo están accesibles y disponibles en formato abierto. formato abierto.*

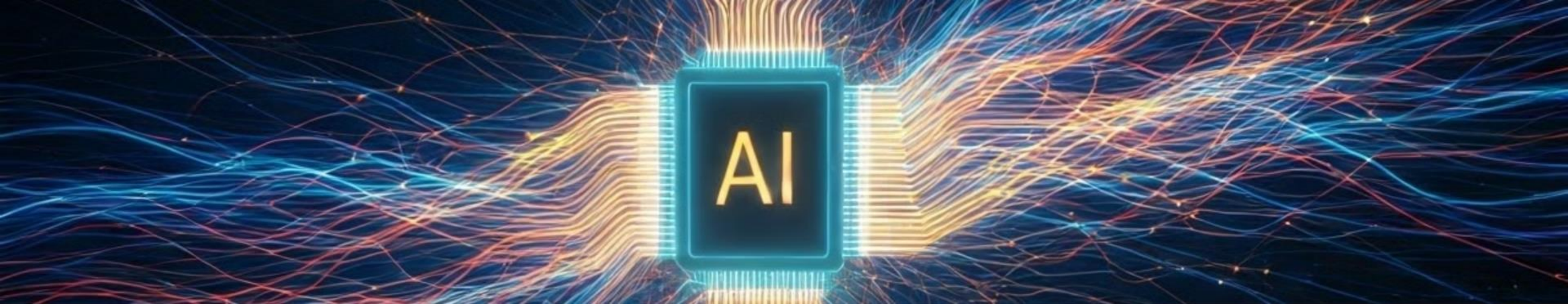
Evaluación de LLM



huggingface.co

🤗 **Open LLM Leaderboa...**

Evaluation of open source Large Language Models (LLMs)



Ventajas de los modelos grandes

Coherencia Mejorada

Mayor coherencia en respuestas largas y mejor comprensión del texto.

Detección de Ambigüedades

Mejor capacidad para identificar y resolver resolver ambigüedades en el lenguaje.

Pertinencia de Información

Habilidad superior para proporcionar información relevante y contextualizada.

Inconvenientes de los LLM

Costes elevados de servicios en la nube

Necesidad de hardware muy potente

Alto consumo de energía

Velocidad de inferencia limitada

Resumen

Esta presentación proporciona una introducción a los modelos de lenguaje (LLM), su funcionamiento, ventajas y desventajas. Abarca desde sus Abarca desde sus fundamentos hasta su uso en diferentes campos.

Explora las arquitecturas de LLM, su entrenamiento, y las diversas aplicaciones que permiten, como la traducción automática, generación de texto y respuesta a preguntas.

Se analizan los diferentes tipos de LLM, desde modelos básicos hasta modelos conversacionales y específicos de dominio.

Estos modelos se basan en redes neuronales profundas, que imitan el funcionamiento del cerebro humano al procesar y aprender de los datos. Mediante técnicas como el aprendizaje profundo, los modelos de lenguaje pueden capturar las sutilezas y complejidades del lenguaje, lo que les permite generar texto coherente y contextualmente relevante.