LARGE LANGUAGE MODELS

Large Language Models

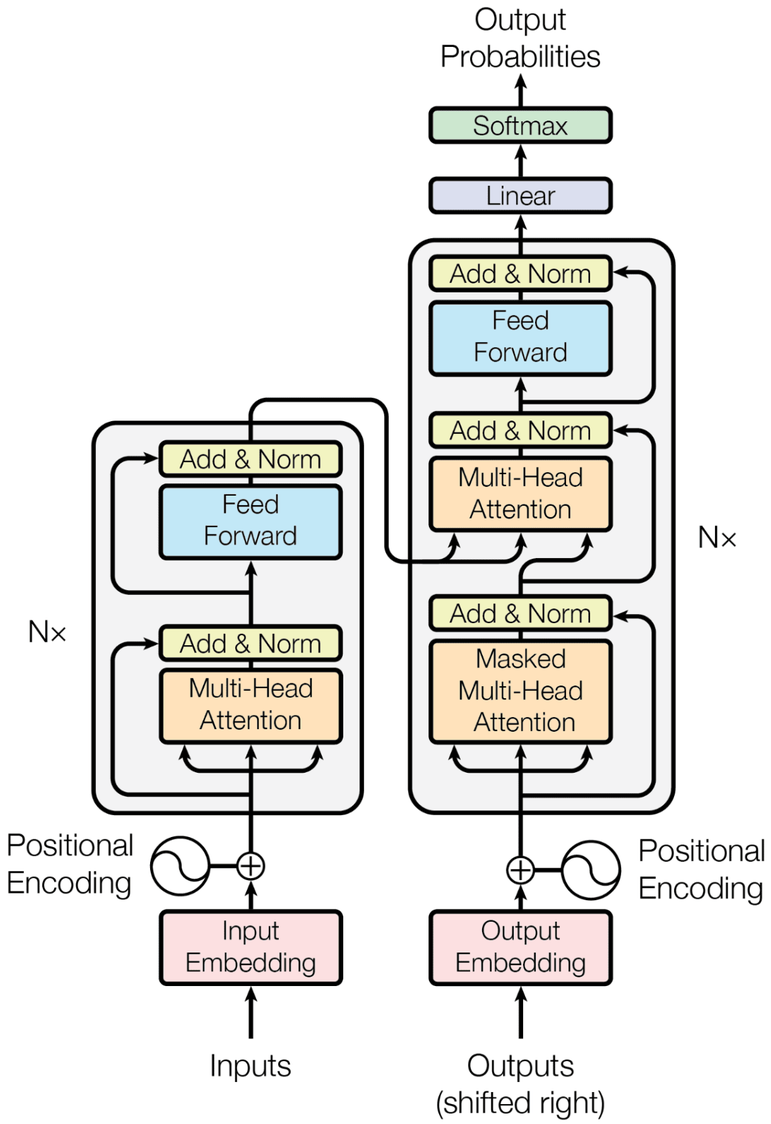
Los modelos de lenguaje han evolucionado enormemente a lo largo de los años, desde sistemas simples hasta los actuales Large Language Models (LLMs).

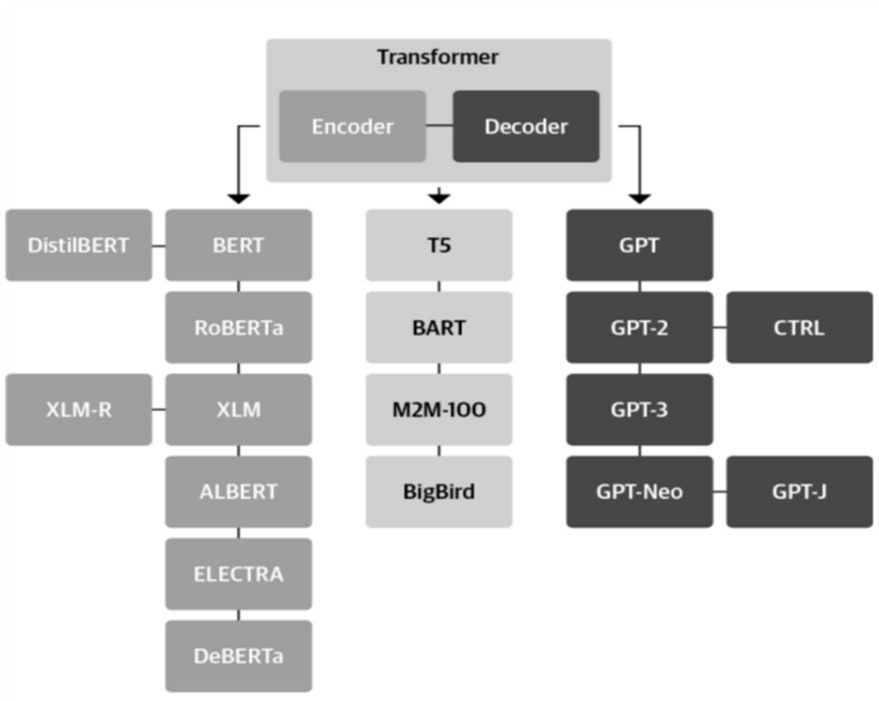
En sus inicios, los sistemas de procesamiento de lenguaje natural se basaban en reglas y patrones gramaticales predefinidos. Eran limitados en su capacidad para comprender el contexto y la ambigüedad del lenguaje humano.

Large Language Models

Década de 2010: Con la llegada de las RNNs, los modelos de lenguaje comenzaron a capturar dependencias a largo plazo en el texto. Sin embargo, aún enfrentaban desafíos en la gestión de secuencias largas y la generación de texto coherente.

2017: La introducción de los Transformers marcó un punto de inflexión en la historia de los modelos de lenguaje. Estas arquitecturas permitieron la atención multi-cabeza y el procesamiento paralelo de secuencias largas, mejorando significativamente la calidad de la generación de texto.


Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)


Transformer
Encoder
Decoder
DistilBERT
BERT
RoBERTa
XLM-R
XLM
ALBERT
ELECTRA
DeBERTa
T5
BART
M2M-100
BigBird
GPT
GPT-2
CTRL
GPT-3
GPT-Neo
GPT-J

Large Language Models

2018 - hoy: Los LLMs hacen uso intensivo de la arquitectura de Transformer y grandes cantidades de datos para lograr un entendimiento y generación de lenguaje natural a escala nunca antes vista.

- **Escalabilidad:** Los LLMs se escalan a decenas o cientos de miles de millones de parámetros.
- **Generalización:** Tienen la capacidad de comprender y generar contenido en varios idiomas y dominios temáticos.
- **Adaptabilidad:** Se pueden ajustar para tareas específicas, como traducción, resumen de texto, chatbots, entre otros.

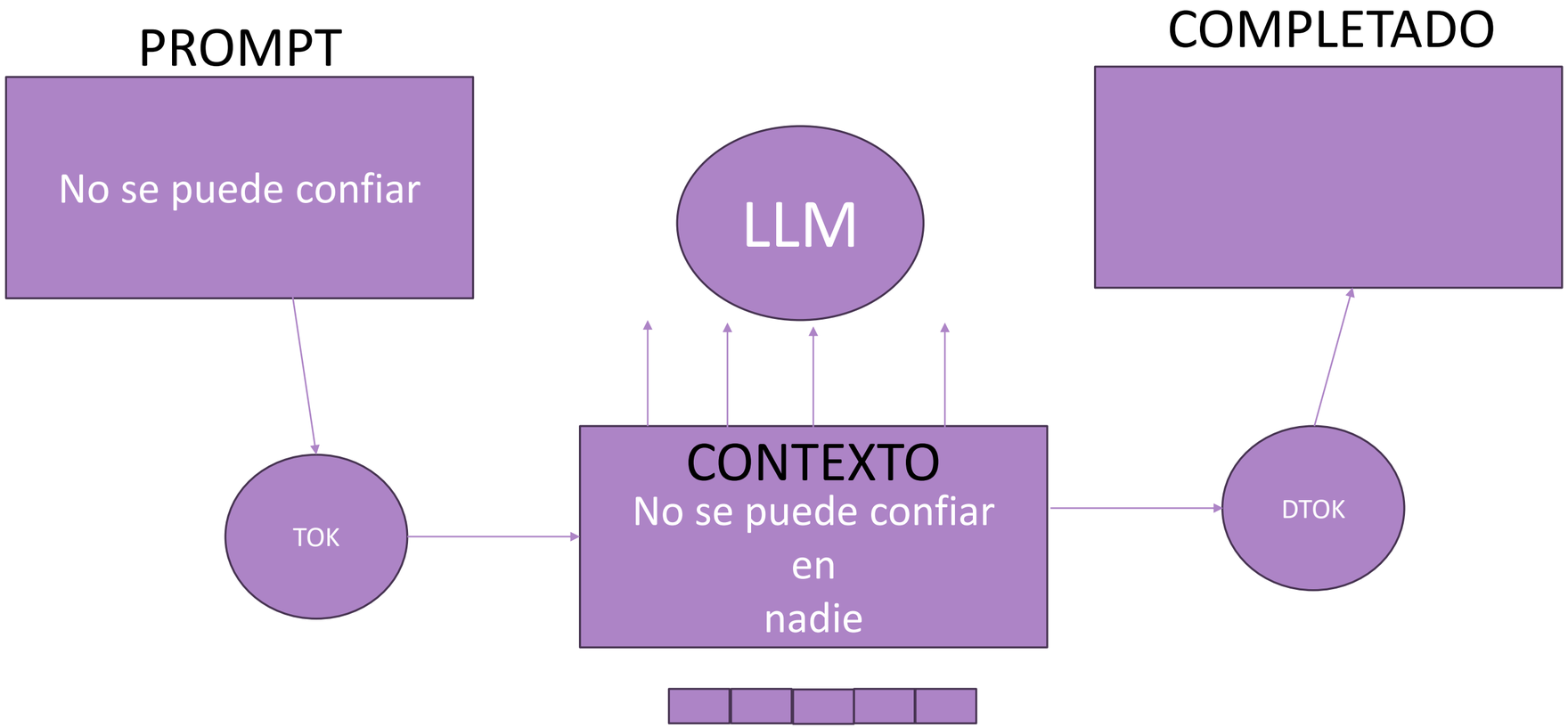
La evolución de los LLMs sigue avanzando, con modelos aún más grandes y sofisticados en desarrollo

Como funcionan los Large Language Models

•Los Large Language Models (LLMs) se basan en arquitecturas de redes neuronales profundas, siendo una de las más destacadas y utilizadas la arquitectura "**Transformer**".

•Estas redes están compuestas por **múltiples capas de atención y feedforward**, lo que les permite comprender las relaciones entre las palabras y los tokens en un contexto amplio.

Como funcionan los Large Language Models



Como funcionan los Large Language Models

•Los LLMs son entrenados en **conjuntos de datos masivos** que contienen una amplia variedad de texto, que van desde páginas web completas hasta libros, artículos y conversaciones en línea.

•Utilizan un proceso de aprendizaje supervisado donde se ajustan los pesos y conexiones de la red para que sea capaz de predecir la siguiente palabra o token en una secuencia de texto basándose en el contexto proporcionado por las palabras previas.

•A través de numerosas iteraciones de entrenamiento en estos enormes conjuntos de datos, los modelos aprenden a **capturar patrones complejos y a entender la gramática, la semántica y el contexto.**

Como funcionan los Large Language Models

Generación de texto: Basados en contexto y patrones estadísticos

•Los LLMs generan texto utilizando un enfoque basado en la probabilidad y el contexto. Dado un **contexto inicial** o una "semilla", el modelo utiliza su comprensión del **lenguaje natural** para predecir la siguiente palabra o token más probable en función de las estadísticas y patrones que ha aprendido durante el entrenamiento.

•A medida que se generan más palabras, el modelo ajusta su predicción en función del contexto acumulado, lo que permite la generación de texto coherente y contextualmente relevante.

•Los LLMs también pueden ser alimentados con **instrucciones específicas** o indicadores de estilo para generar contenido que se ajuste a ciertos criterios, como escribir en un tono formal o informal.

LLM Development (using pre-trained APIs)

- NO ML expertise needed
- NO training examples
- NO need to train a model
- Thinks about prompt design

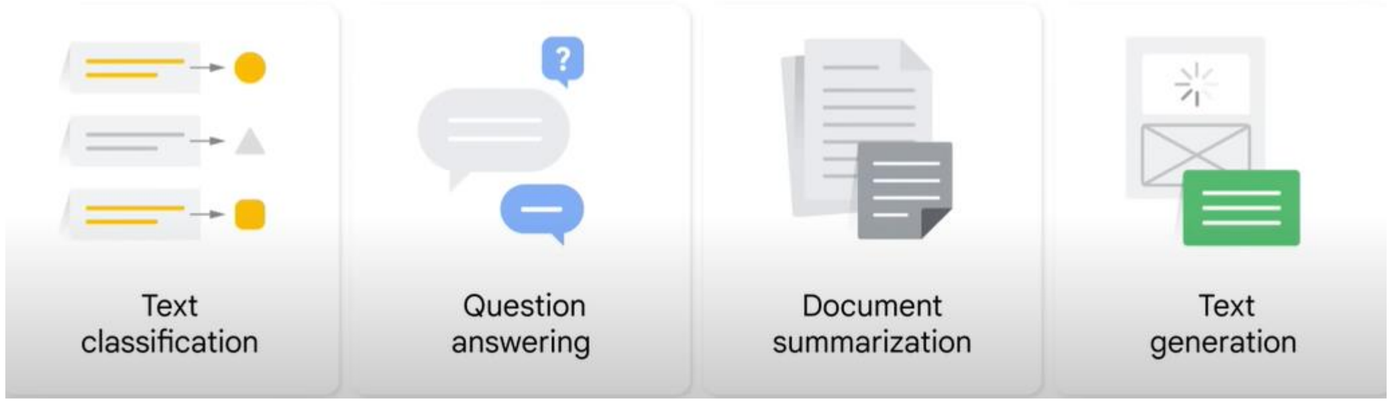
Traditional ML Development

- YES ML expertise needed
- YES training examples
- YES need to train a model
- YES compute time + + hardware
- Thinks about minimizing a loss function

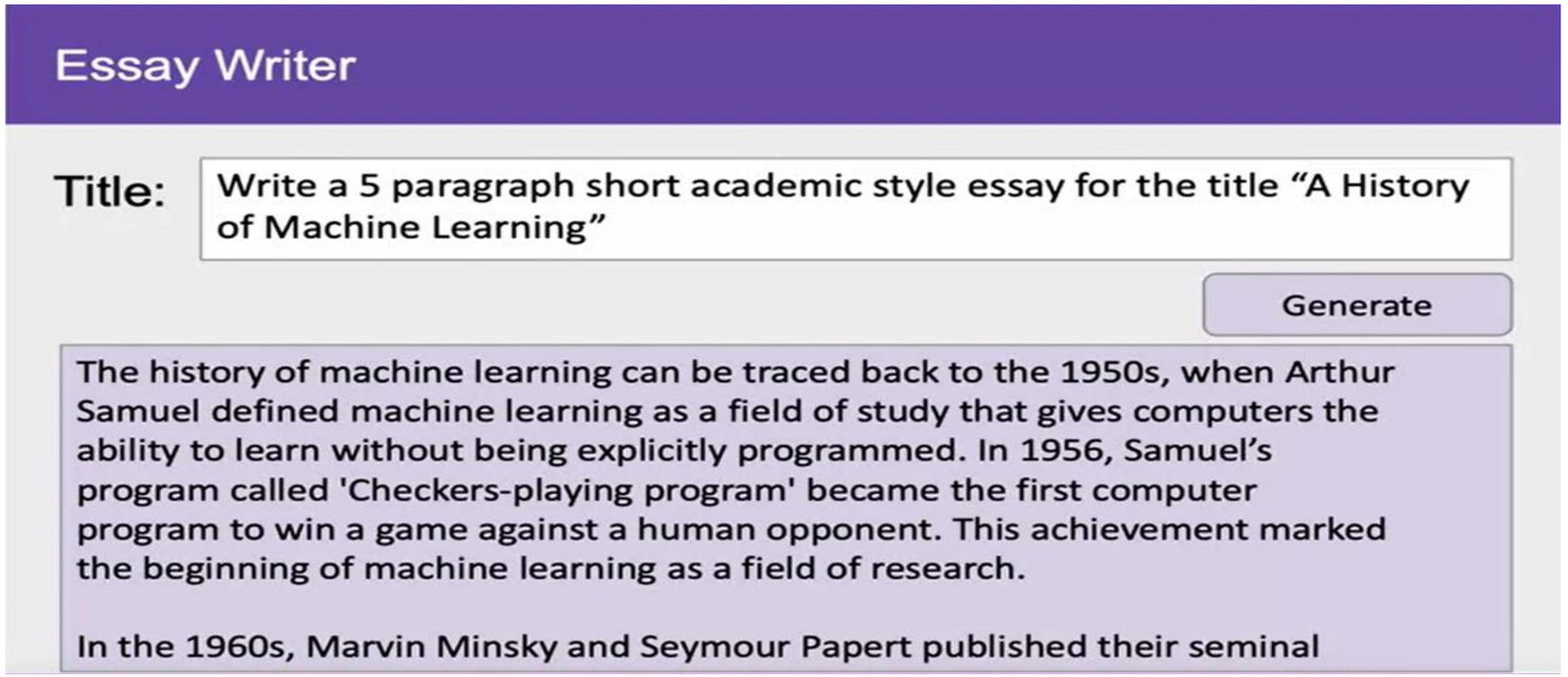
Los LLMs pueden aplicarse en tareas de:

- Traducción automática
- Resumen de documentos
- Respuesta a preguntas
- Chatbots
- Extracción de entidades
- Generación de código

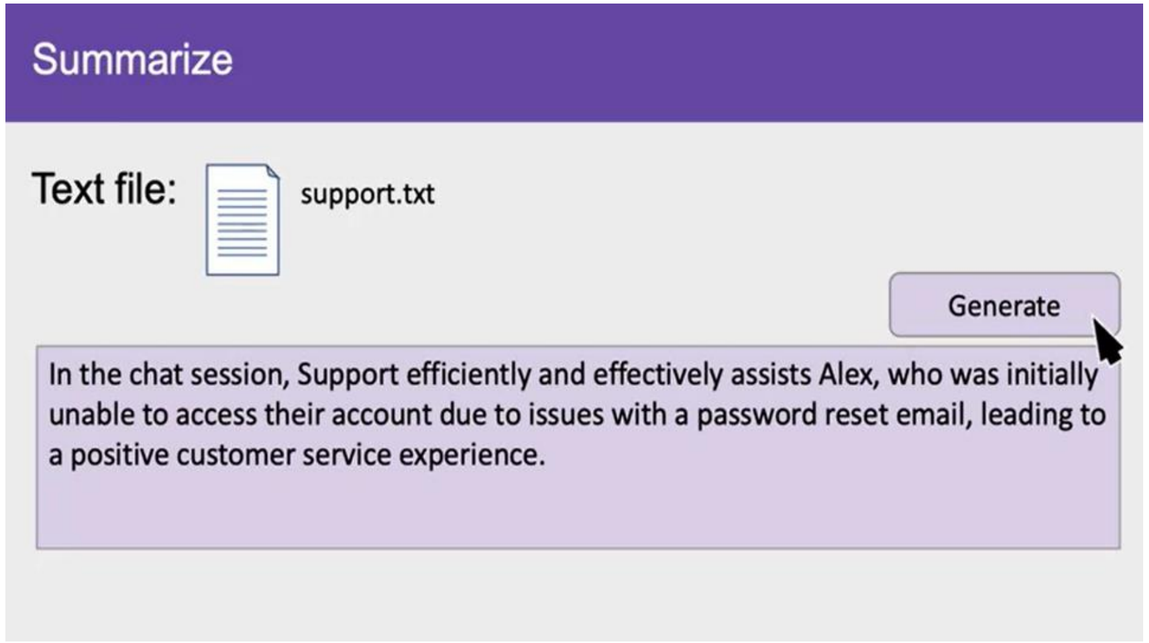
•Su capacidad para generar texto coherente y contextualmente relevante los hace valiosos en una amplia variedad de aplicaciones en incontables campos



- Generación de texto

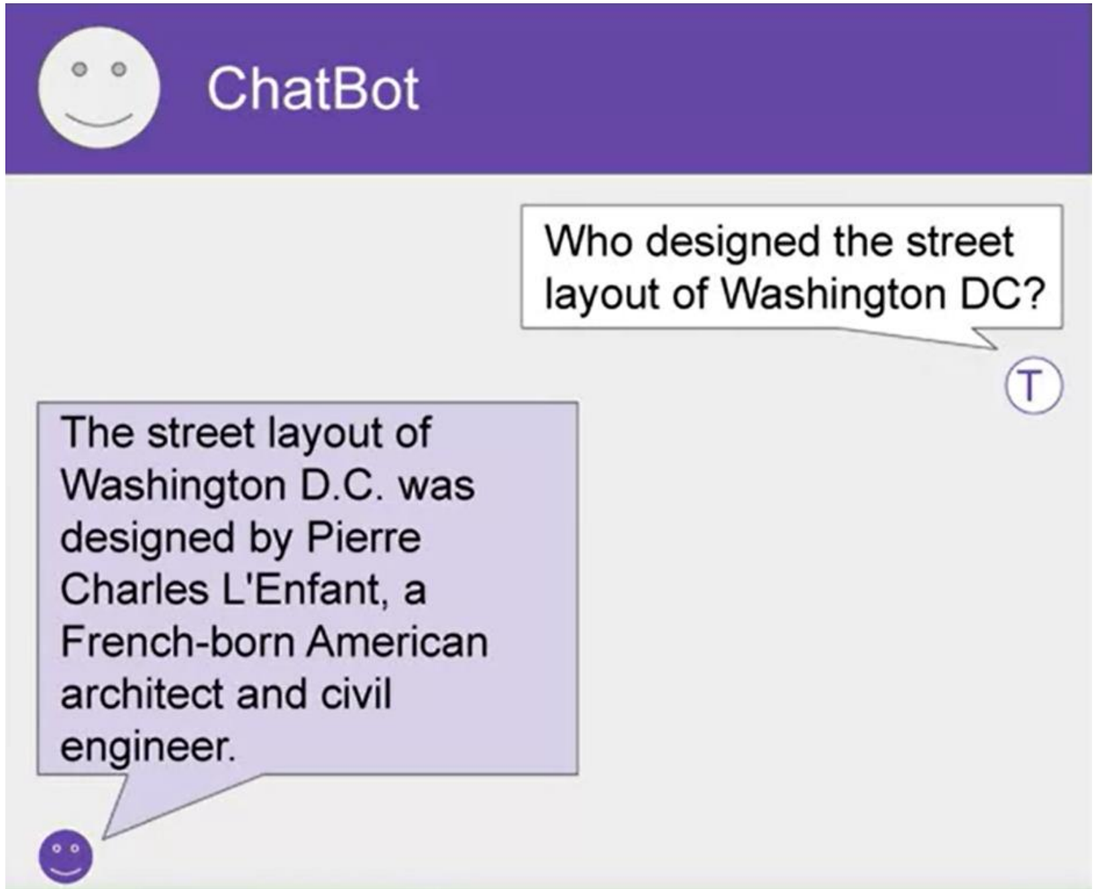


- Generación de resúmenes

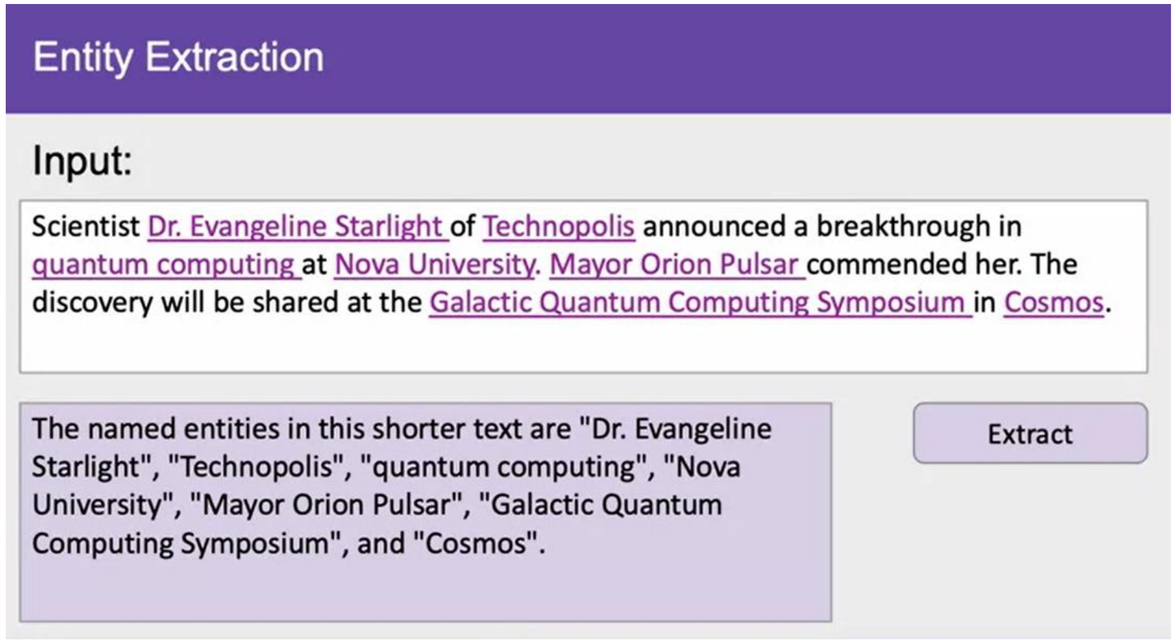


- Traducción automática

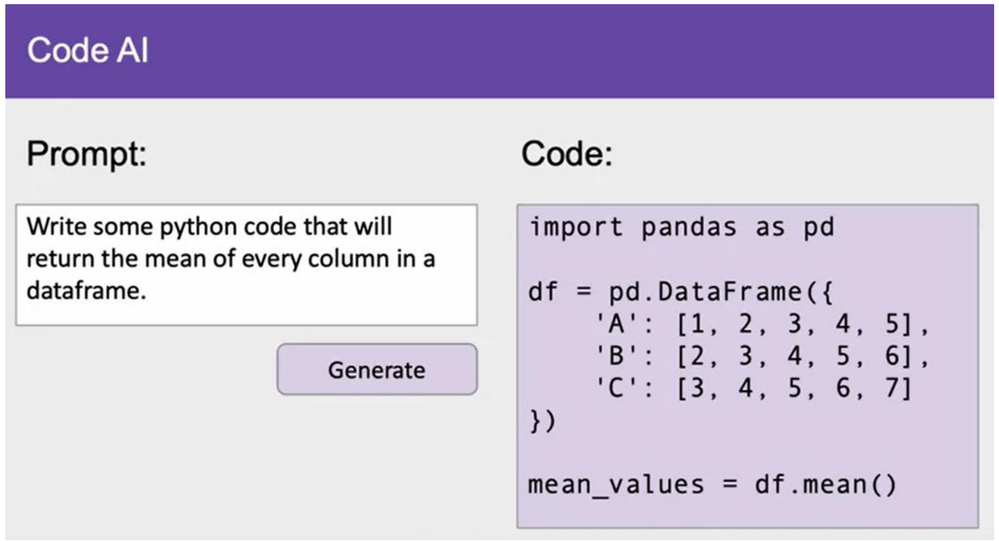
- Chatbots


ChatBot
Who designed the street layout of Washington DC?
T
The street layout of Washington D.C. was designed by Pierre Charles L'Enfant, a French-born American architect and civil engineer.

- Extracción de entidades



- Generación de código



Que LLMs tenemos hasta ahora

OPEN SOURCE MODELS

CLOSE SOURCE MODELS

OPEN SOURCE MODELS

https://huggingface.co/models

https://huggingface.co/HuggingFaceH4

https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

Modelos de código abierto

Un LLM de código abierto es un tipo de LLM cuyo código fuente está disponible públicamente, permitiendo a los usuarios acceder y modificar los pesos del modelo.

Los LLMs de código abierto generalmente son gratis y pueden ser entrenados para determinadas tareas.

Falcon

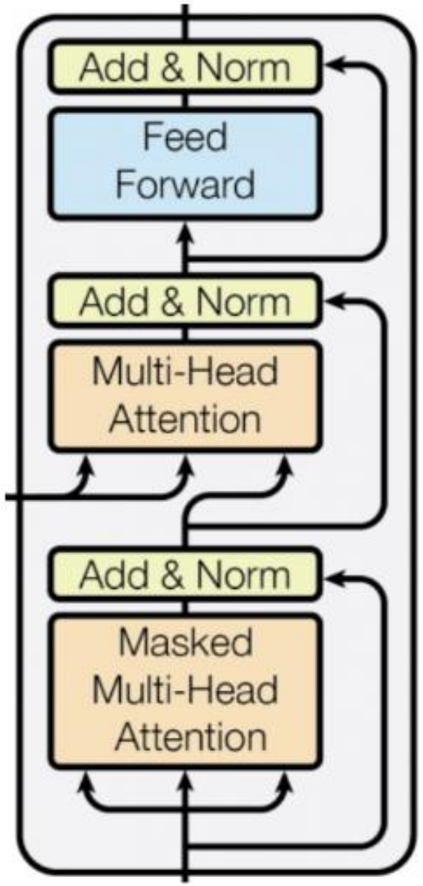
Falcon

- Falcon es un LLM generativo desarrollado por el Technology Innovation Institute
- La familia de LLMs Falcon está compuesta por tres modelos: Falcon-7B, Falcon-40B, y Falcon-180B
- El modelo está basado en el conjunto de datos RefinedWeb, descrito en el trabajo de investigación titulado "The Refined Web Dataset for Falcon LLM"

Arquitectura

- Falcon es un modelo de lenguaje Transformer con solo un descodificador.

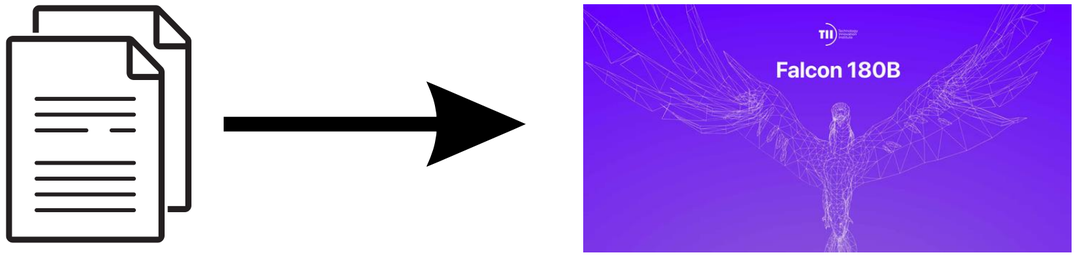
- Esta arquitectura difiere de otros modelos basados en la estructura transformer, como T5, que emplean tanto capas de codificación como de descodificación.



Decoder layer

Puntos clave

- Falcon 180B fue lanzado el 6 de septiembre de 2023
- En un LLM con 180 billones de parámetros, lo que le convierte en uno de los modelos de código abierto más grandes disponible.
- El LLM Falcon 180B ha sido fine-tuned en 3.5 trillones de tokens de texto.



Puntos clave

- Falcon 180B es compatible con los principales idiomas, incluyendo inglés, alemán, español y francés.
- Tiene capacidades limitadas en italiano, portugués y árabe, con planes de ampliar el soporte de idiomas en el futuro.

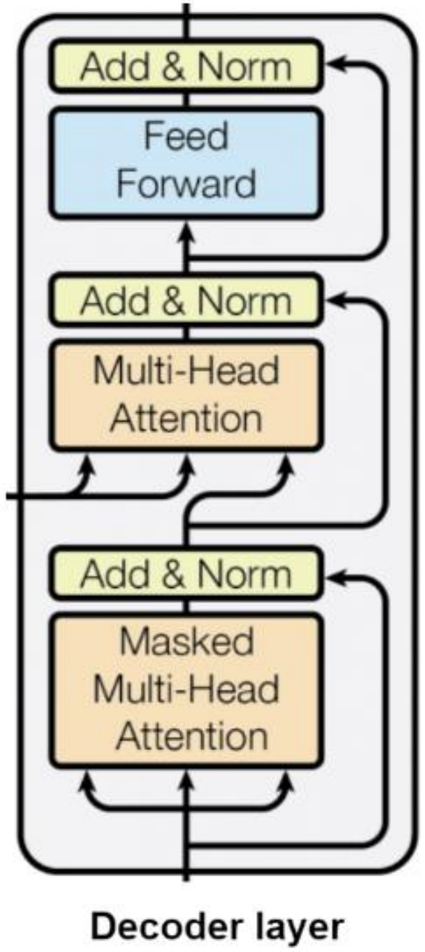
Llama2

Llama2

- LLaMA 2 es la siguiente generación de la familia de LLMs LLaMA, desarrollada y lanzada por Meta AI
- Fue lanzada por Meta AI en julio de 2023.
- Llama 2 es una colección de modelos generativos de texto pre-entrenados y fine-tuned de entre 7 y 10 billones de parametros.

Arquitectura

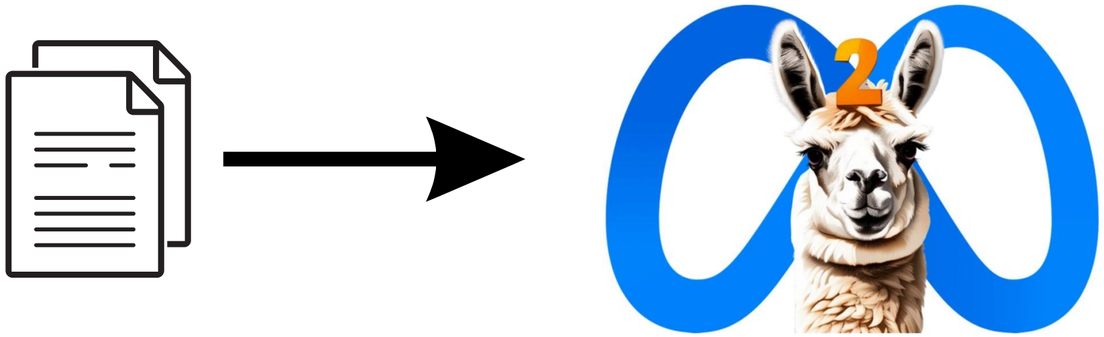
Llama2 es también un modelo de lenguaje basado en la estructura Transformer con solo un decodificador.



Decoder layer

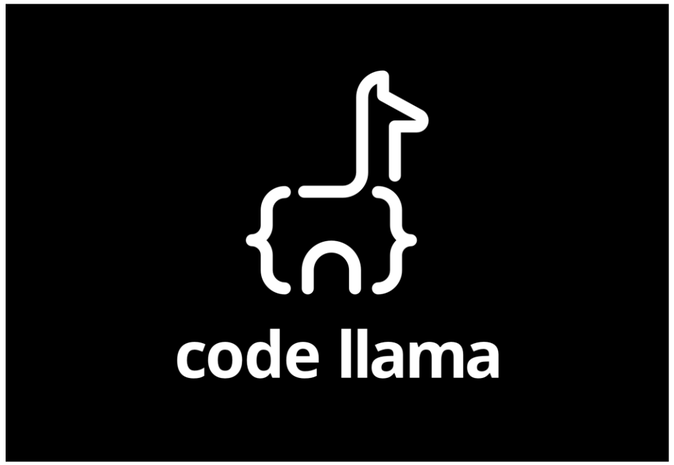
Puntos clave

- LLaMA 2 70B fue entrenado en un conjunto de 2 trillones de tokens de datos de fuentes disponibles públicamente.
- El trabajo de LLaMA 2 se titula "Llama 2: Open Foundation and Fine-Tuned Chat Models" y fue publicado en arXiv el 18 de Julio de 2023.



Code llama

- Code Llama está formado a partir de Llama2 y se emplea para generar fragmentos de código, sugerir la finalización de código o depurar código.
- Code Llama ha sido fine-tuned en un gran conjunto de datos para mejorar su rendimiento.



Code llama

Code Llama posee tres versiones:

- Code Llama: un modelo general para tareas de codificación.
- Code Llama - Python: adaptado para tareas en Python.
- Code Llama - Instruct: diseñado para seguir y ejecutar instrucciones de lenguaje natural.

BLOOM

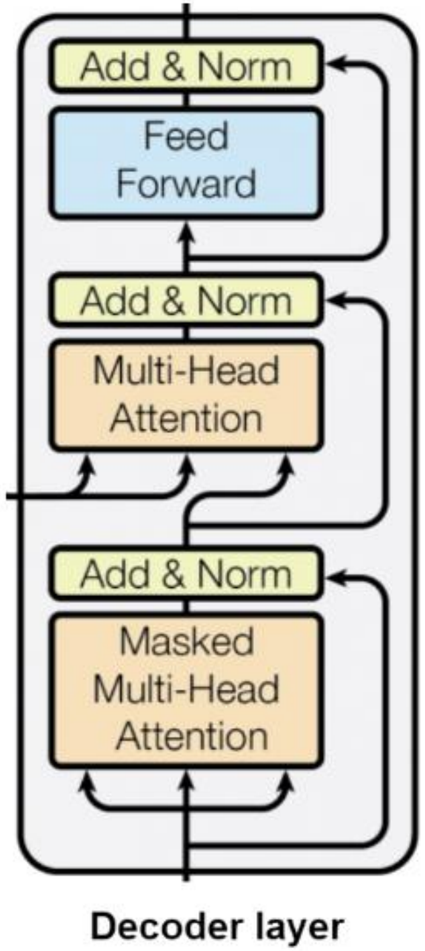
Bloom

- BLOOM es un modelo abierto de lenguaje multilingüe. Ha sido desarrollado por más de 1000 investigadores de IA de más de 70 países y más de 250 instituciones.

- El proyecto BLOOM involucró a seis principales equipos: BigScience de HuggingFace, DeepSpeed de Microsoft, Megatron-LM de NVIDIA, IDRIS/GENCI, PyTorch y los voluntarios del grupo de trabajo de ingienería de BigScience.

Arquitectura

Bloom también es un modelo de lenguaje basado en la estructura Transformer con solo un decodificador.



Decoder layer

Puntos clave

- BLOOM ha sido entrenado durante 3.5 meses en 384 A100-80GB GPUs.
- BLOOM tiene 176 billones de parámetros, convirtiéndolo en uno de los LLMs de mayor tamaño.

Puntos clave

- BLOOM puede usarse para diferentes tareas NLP como la generación de textos, el resumen de textos, embeddings, clasificación, búsqueda semántica y traducción.

- El trabajo de investigación original sobre BLOOM se titula: "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model" y fue publicado en arXiv en noviembre de 2022

T5

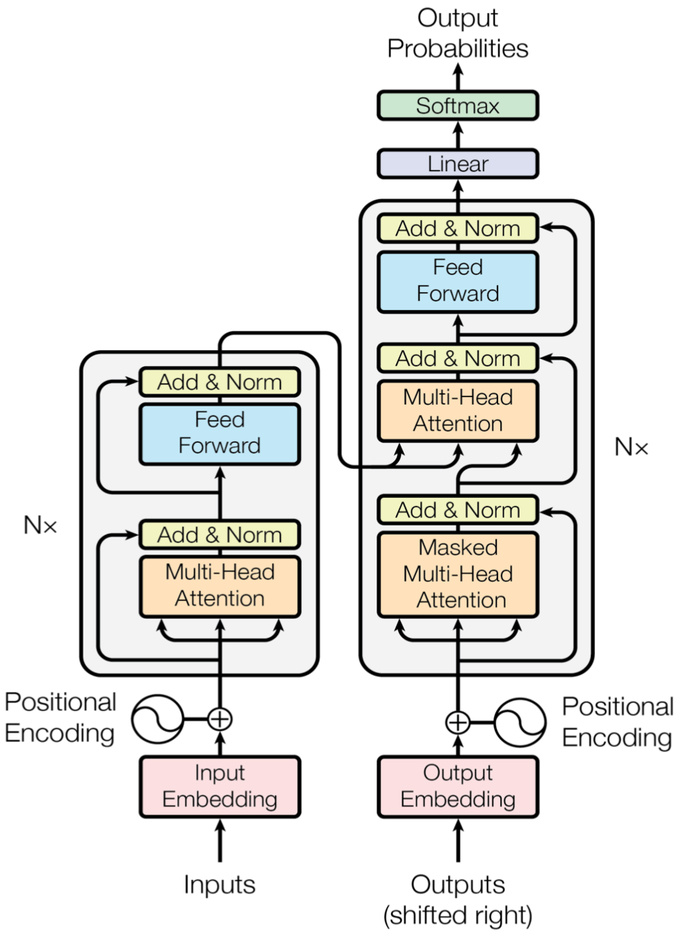
T5

- T5 es un modelo de código abierto creado por el equipo de IA de Google que puede ser fine-tuned en una amplia gama de tareas NLP.

- El trabajo de investigación original del LLM T5 se titula: "Explorting the Limits of Transfer Learning with a Unified Text-to-Text Transformer" y fue publicado en arXiv en 2019.

Architecture

- El modelo T5 está basado en la arquitectura estándar codificador-decodificador, similar a la estructura Transformer propuesta en Vaswani et al.
- La arquitectura del modelo T5 consiste en una pila de bloques transformer tanto en el codificador como en el decodificador.



Key points

El número de parámetros del LLM T5 (modelo de lenguaje transformer text-to-text) varía según la versión específica del modelo. Algunas estimaciones de la cantidad del número de parámetros para diferentes modelos T5 son:

- T5-Small: 60 millones de parámetros
- T5-Base: 220 millones parámetros
- T5-Large: 770 millones parámetros
- T5-3B: 3 billones de parámetros
- T5-11B: 11 billones de parámetros

Flan T5

- Flan-T5 es una versión mejorada del modelo de lenguaje T5 que ha sido fine-tuned en una combinación de tareas. Incluye las mismas mejoras que la versión 1.1 del modelo T5 y puede ser empleado directamente sin fine-tuning.
- Flan-T5 ha sido fine-tuned en más de 1000 tareas adicionales abarcando más idiomas que T5.

MODELOS DE CÓDIGO PRIVADO

Modelos de código privado

Los LLMs de código cerrado son aquello cuyo código no está disponible públicamente. Normalmente son desarrollados por grandes compañías

GPT4

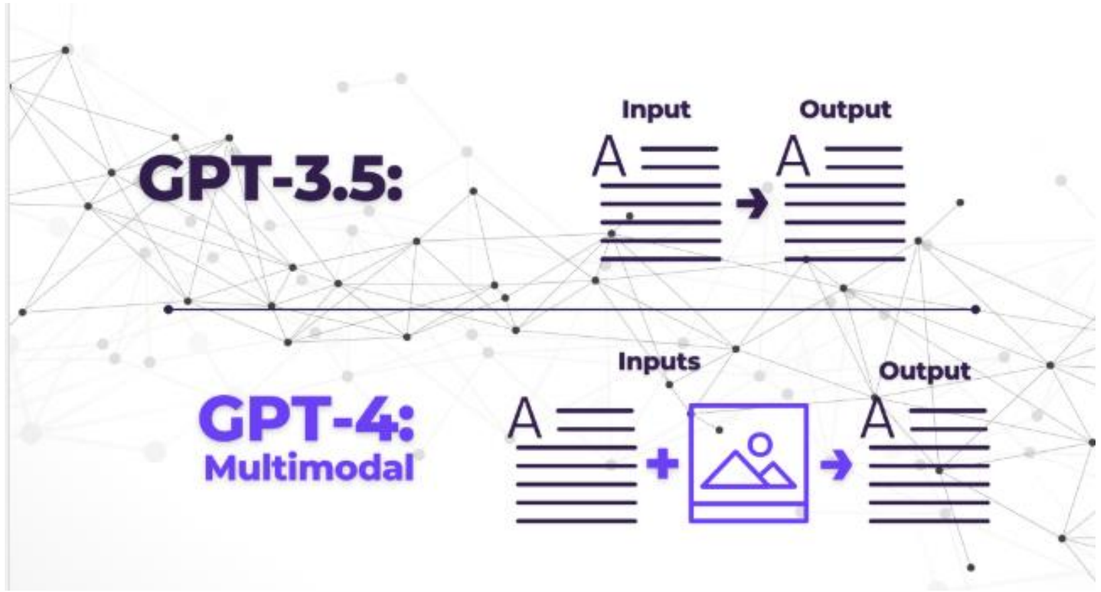
GPT4

- GPT-4 es un gran modelo de lenguaje multimodal creado por OpenAI y es el cuarto de su serie de modelos base GPT.

- GPT-4 fue lanzado oficialmente el 14 de marzo de 2023.

GPT4

- GPT-4 puede analizar y comentar imágenes y gráficos, además de texto.
- GPT-4 tiene un límite de 32.000 tokens como máximo, lo cual supone un gran aumento con respecto a los 4.000 tokens de GPT-3



What is funny about this image? Describe it panel by panel.

Source: hmmm (Reddit)

GPT-4

The image shows a package for a "Lightning Cable" adapter with three panels.

Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port.

Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it.

Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end.

The humor in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port.

Palm2

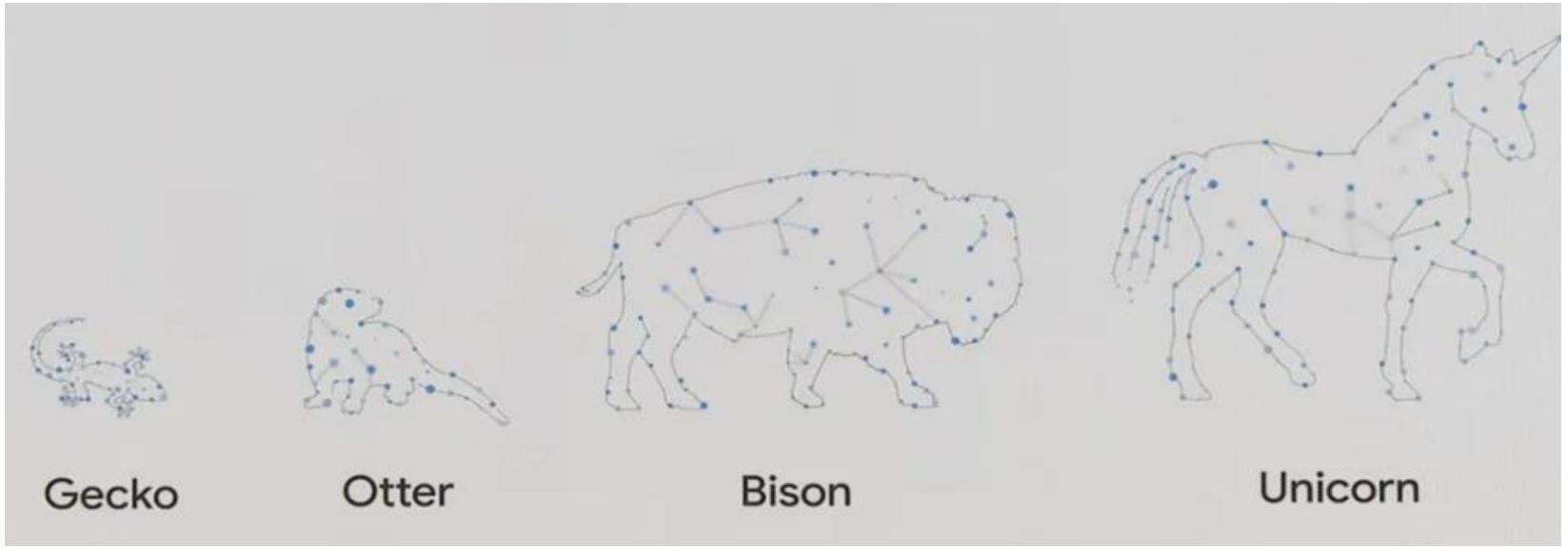
Palm2

- PaLM 2 es la siguiente generación de LLMs de Google que se basa en el legado de investigación innovadora de Google en aprendizaje automático e IA responsable.

- Destaca en tareas de razonamiento avanzado, incluyendo código y matemáticas, clasificación y respuesta de preguntas, traducción y dominio multilingüe, y generación de lenguaje natural.

Palm2 Models

PaLM 2 está disponible en cuatro versiones según el tamaño. De menor a mayor: Gecko, Otter, Bison, and Unicorn,



Key points

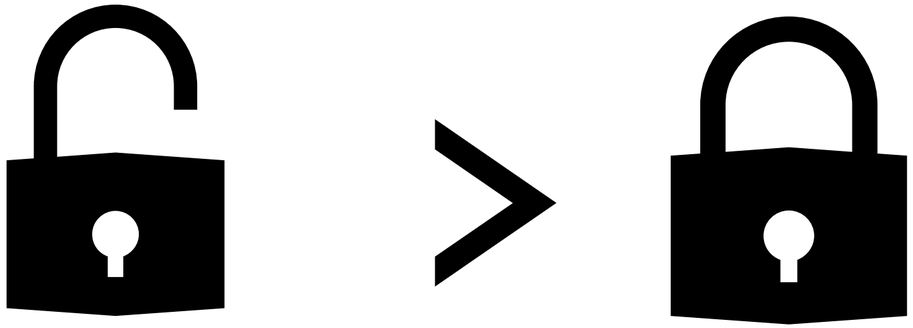
- PaLM 2 está entrenado en 540 billones de parámetros.
- Google provee acceso a PaLM 2 a través de la API The Vertex PALM (un servicio de Google Cloud que permite el uso y entrenamiento de modelos generativos).

Open LLMS vs closed LLMS

Open LLMS vs closed LLMS

- Los LLMs de código abierto pueden ser personalizado para satisfacer necesidades específicas, añadiendo características concretas o entrenando en conjuntos de datos específicos.

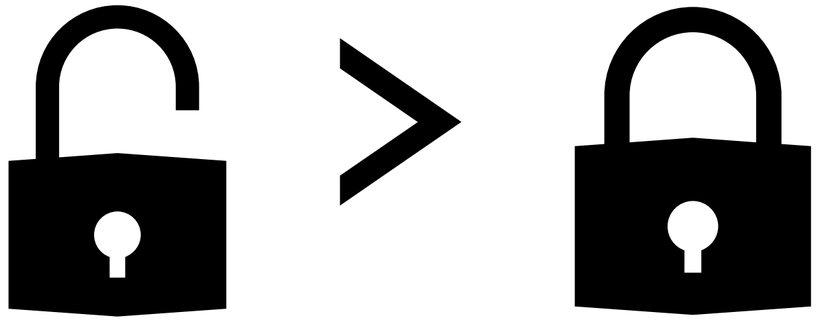
- Los LLMs de código abierto son típicamente menos costosos que LLMs propietarios ya que no poseen tarifas de licencia.



Open LLMS vs closed LLMS

- Los LLMs de código abierto fomentan una cultura de innovación y colaboración, ya que los usuarios pueden contribuir al desarrollo y mejora de los modelos.

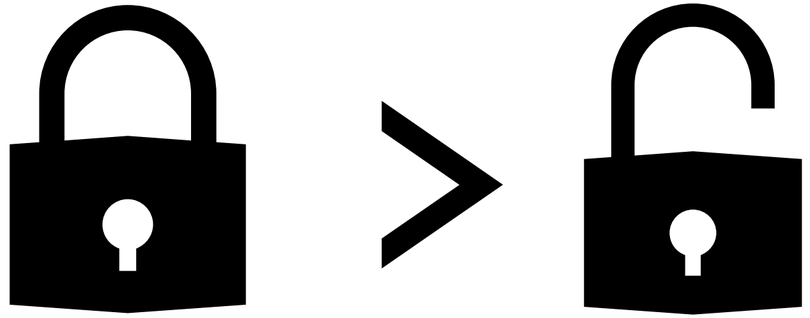
- Los LLMs de código abierto ofrecen seguridad y privacidad mejoradas ya que el código fuente está disponible para su revisión por la comunidad.



Open LLMS vs closed LLMS

- Los LLMs de cógido privado a menudo prometen un rendimiento consistente y de gran calidad, ya que son desarrollados por grandes compañias con los medios económicos necesarios para respaldar su investigación y desarrollo.

- Los LLMs de código privado pueden ser más fáciles de usar ya que suelen ser desarrollados para un perfil de usuario específico y pueden poseer por tanto interfaces más intuitivas.



LAB