

# IA para la DOCENCIA

Otoño Tech-UPM



## Taller 1: ¿Cómo funciona la IA generativa?: Fundamentos de IA y LLMs

# ¿CÓMO FUNCIONA LA IA GENERATIVA?

## Fundamentos de la IA y LLMs

Raúl Lara Cabrera (raul.lara@upm.es)

Fernando Ortega Requena (fernando.ortega@upm.es)



¿Qué es la  
Inteligencia Artificial?

# ¿QUÉ ES LA INTELIGENCIA ARTIFICIAL?

La inteligencia artificial (IA) es una rama de la **informática** que se dedica a crear sistemas y tecnologías capaces de realizar tareas que normalmente requieren **inteligencia humana**. Estas tareas incluyen cosas como entender el lenguaje natural, reconocer patrones en imágenes, tomar decisiones, resolver problemas y **aprender** de la experiencia.

La IA se basa en el uso de **algoritmos** y modelos matemáticos que permiten a las máquinas "aprender" a partir de datos y mejorar su desempeño con el tiempo. El objetivo principal es desarrollar sistemas que puedan imitar el pensamiento humano y realizar tareas de manera autónoma, sin intervención humana directa.

# LOS PRIMEROS DÍAS (1950S - 1960S)

Creación del concepto de la "máquina inteligente".



Alan Turing propone el Test de Turing en 1950.



El término "Inteligencia Artificial" es acuñado en la conferencia de Dartmouth en 1956.

# DESARROLLO DE REDES NEURONALES (1980S)

Desarrollo de redes neuronales artificiales.



Introducción de algoritmos de aprendizaje supervisado.



Las redes neuronales comenzaron a ganar terreno en tareas específicas como el reconocimiento de patrones.

# ERA DEL APRENDIZAJE AUTOMÁTICO (2000S)

Algoritmos de aprendizaje automático (Machine Learning)



Predicción de datos, motores de recomendación y análisis predictivo



El uso de grandes volúmenes de datos mejoró la precisión y efectividad de los modelos

# REVOLUCIÓN DEL APRENDIZAJE PROFUNDO (2010S)

Expansión del aprendizaje profundo (Deep Learning)



Avances en redes neuronales convolucionales (CNN) y redes recurrentes (RNN)



DeepMind's AlphaGo vence a un campeón mundial en Go en 2016

# EL BOOM DE LA INTELIGENCIA ARTIFICIAL (2020S)



**ChatGPT**

**LLAMA**

**Gemini**



**Claude**



**deepseek**



**Copilot**



¿Cómo funciona la  
Inteligencia Artificial?

# ¿CÓMO FUNCIONA LA INTELIGENCIA ARTIFICIAL?



Datos



Algoritmos



Potencia de cómputo

# PIPELINE DE LA IA

Recopilación de datos

Procesamiento de datos

Algoritmos de aprendizaje

Optimización

Implementación y ajuste

# ¿APRENDE LA IA?



Se dice que la IA aprende, ¿es verdad?



No, la IA optimiza funciones matemáticas.

# OPTIMIZANDO UNA FUNCIÓN

Queremos predecir el precio de venta de una casa

Tenemos:

- Las características de la casa
  - + *Metros cuadrados*
  - + *Número de habitaciones*
  - + *Año de construcción*
  - + *Tipo de vivienda*
  - + ...
- Las ventas de casas de los últimos 10 años.



Definimos un modelo que permita combinar las características de la casa para obtener su precio de venta.



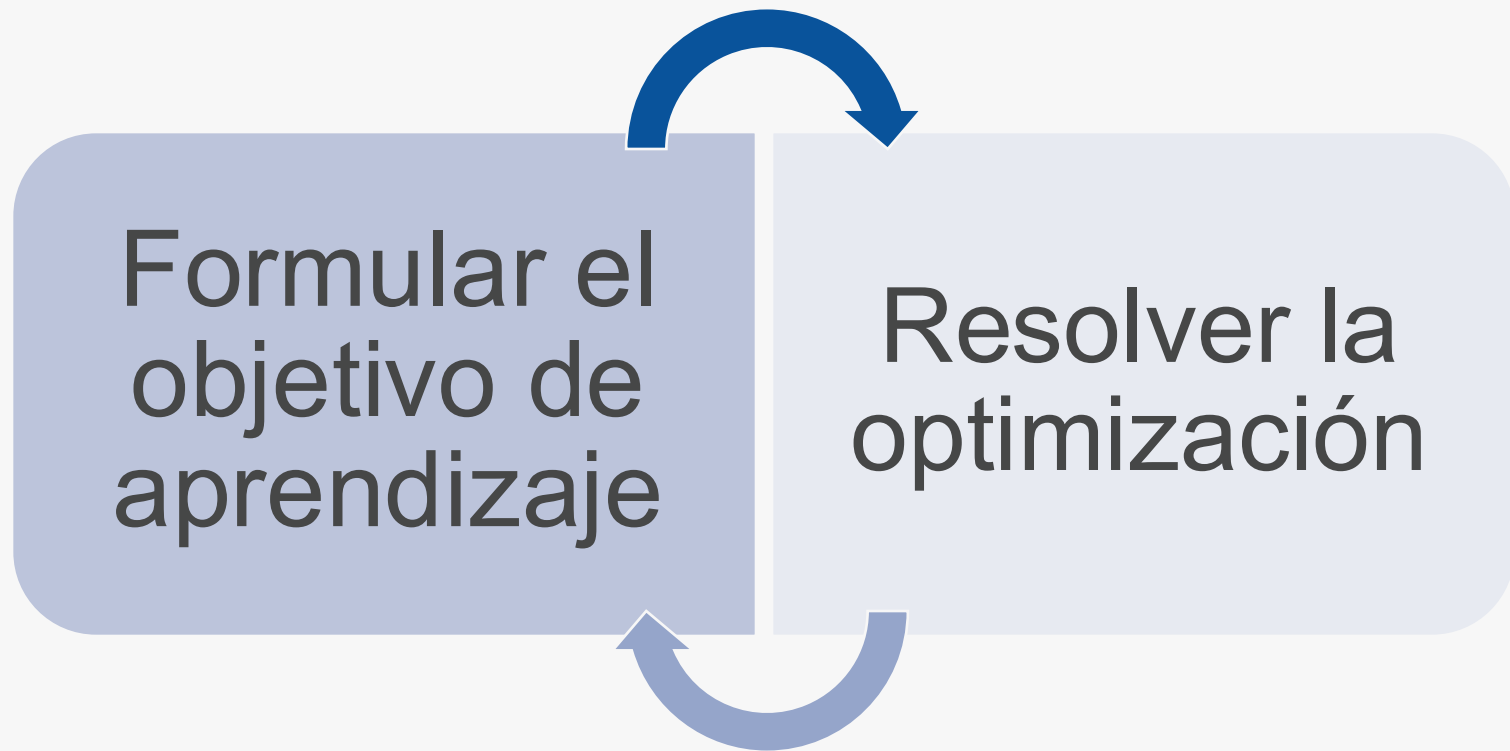
$$\text{precio} = 100000 + 8953 * m^2 + 23764 * n^{\circ}hab + \dots$$

$$\text{precio} = w_0 + w_1 * x_1 + \dots + w_n * x_n$$

Optimizamos:

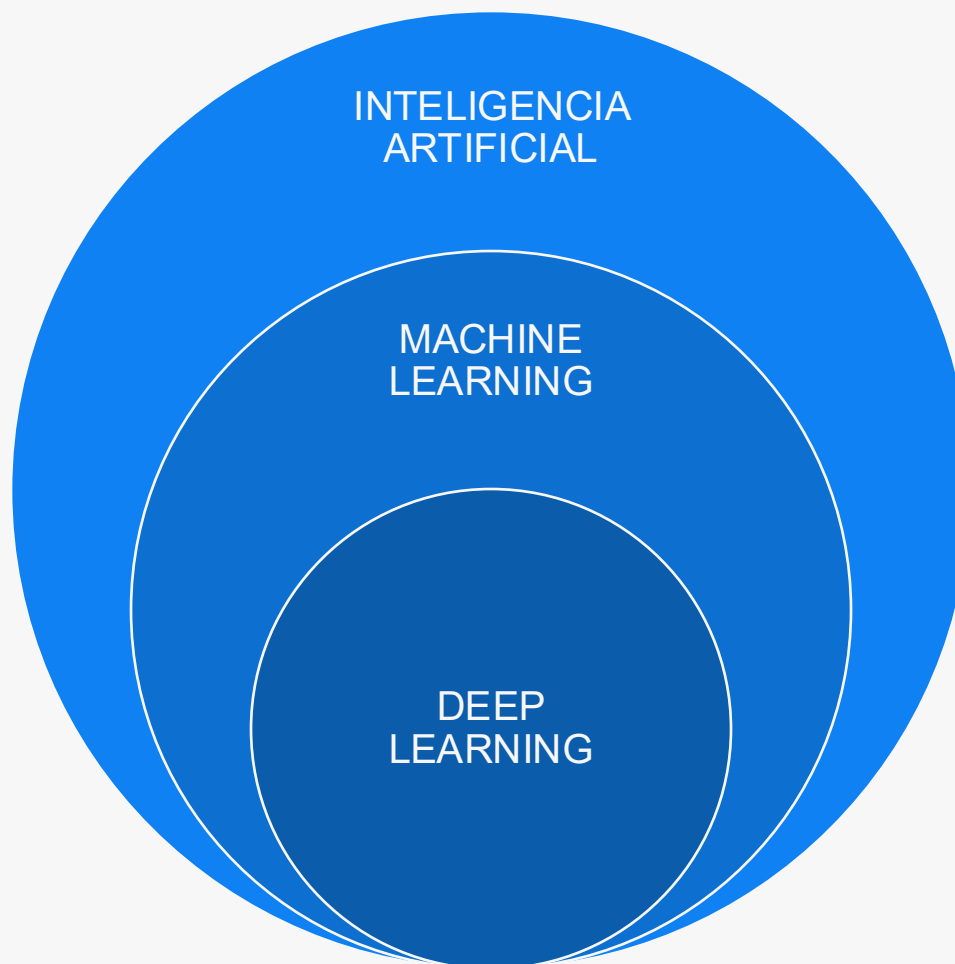
$$\min_w \sum_{\text{casa}} [\text{precio}_{\text{casa}} - (w_0 + w_1 * x_{\text{casa},1} + \dots + w_n * x_{\text{casa},n})]^2$$

# DOS GRANDES DESAFÍOS





¿Qué son las redes  
neuronales?



# DEEP LEARNING VS MACHINE LEARNING

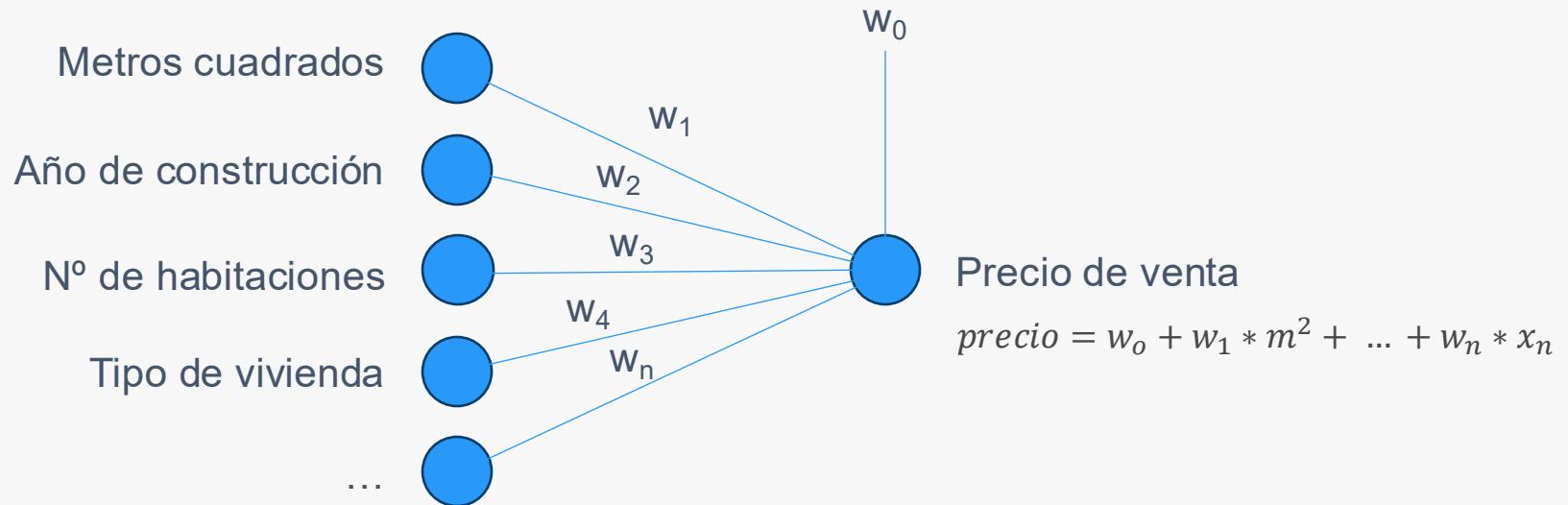
El **deep learning** utiliza redes neuronales artificiales compuestas por capas de neuronas, también conocidas como perceptrones. Estas redes se inspiran en la estructura y funcionamiento del cerebro humano, donde las neuronas están conectadas y se comunican entre sí.

Mientras que el machine learning clásico depende de **técnicas estadísticas** y modelos más “simples”, el deep learning construye modelos matemáticos avanzados mediante la agregación de **millones de parámetros**.

En machine learning tradicional, las características de los datos deben ser diseñadas **manualmente** por expertos, mientras que en deep learning, las redes neuronales pueden aprender **automáticamente** qué características son más relevantes.

El deep learning suele requerir **grandes cantidades** de datos para obtener buenos resultados, mientras que algunos algoritmos de machine learning pueden funcionar bien con conjuntos de datos más pequeños.

# ESQUEMA DE UNA RED NEURONAL

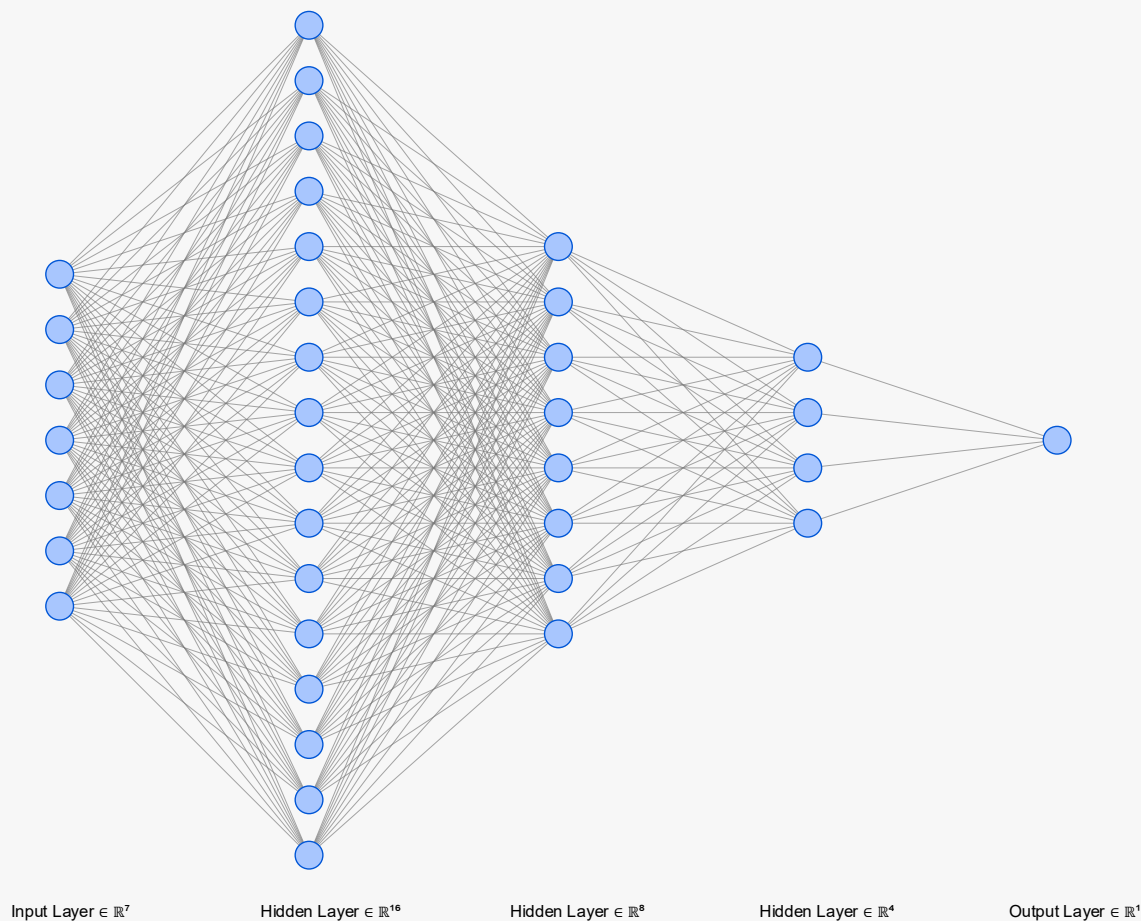


Optimización mediante  
*backpropagation*

*n+1 parámetros a  
aprender*

Funciones de  
activación no lineales

# ESQUEMA DE UNA RED NEURONAL



# REDES NEURONALES FUNCIONANDO

Tinker With a **Neural Network** Right Here in Your Browser.  
Don't Worry, You Can't Break It. We Promise.

Epoch: 000,000 | Learning rate: 0.03 | Activation: Tanh | Regularization: None | Regularization rate: 0 | Problem type: Classification

**DATA**  
Which dataset do you want to use?  
Ratio of training to test data: 50%  
Noise: 0  
Batch size: 10

**FEATURES**  
Which properties do you want to feed in?  
 $X_1$   
 $X_2$   
 $X_1^2$   
 $X_2^2$   
 $X_1 X_2$

**2 HIDDEN LAYERS**  
4 neurons | 2 neurons

**OUTPUT**  
Test loss 0.504  
Training loss 0.497

The outputs are mixed with varying weights, shown by the thickness of the lines.

This is the output from one neuron.

Visita: <https://playground.tensorflow.org/>



Attention Is All  
You Need

# DE RECONOCER PATRONES A ENTENDER PALABRAS

- Las redes neuronales aprender relaciones entre números (vectores)
- Pero el lenguaje son palabras
- Necesitamos representar las palabras como números

Vocabulario:

perro, gato, manzana, pera

One-hot encoding:

perro -> [1,0,0,0]

gato -> [0,1,0,0]

manzana -> [0,0,1,0]

pera -> [0,0,0,1]

Muchos problemas

# DE PALABRAS A ESPACIOS SEMÁNTICOS

- Representación de palabras en espacios semánticos (**embeddings**)
- Word2Vec:
  - Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. *Proceedings of the 1st International Conference on Learning Representations (ICLR 2013)*, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings.
  - **CBOW**: dado un contexto, predice la palabra
  - **Skip-gram**: dada una palabra, predice el contexto

Lenguaje:

El perro y el gato son animales...

Juan merienda manzana o pera...

...

Embedding (tamaño 4):

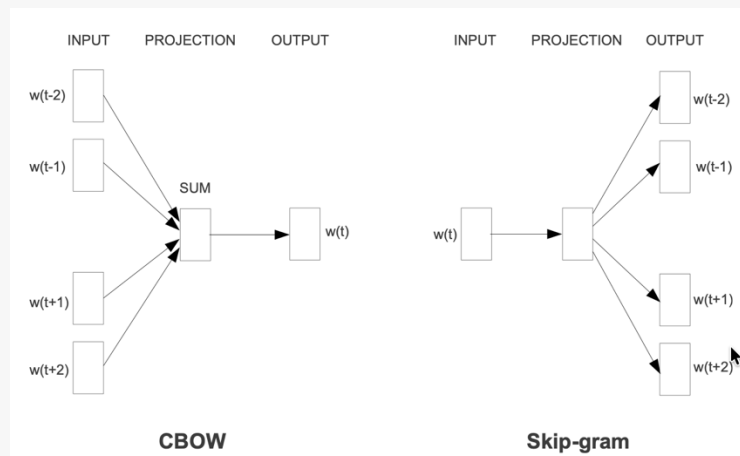
perro -> [0.25, 0.80, 0.10, 0.50]

gato -> [0.20, 0.75, 0.15, 0.45]

manzana -> [0.90, 0.10, 0.85, 0.20]

pera -> [0.85, 0.15, 0.80, 0.25]

...



# EL CONTEXTO IMPORTA

## Redes convolucionales

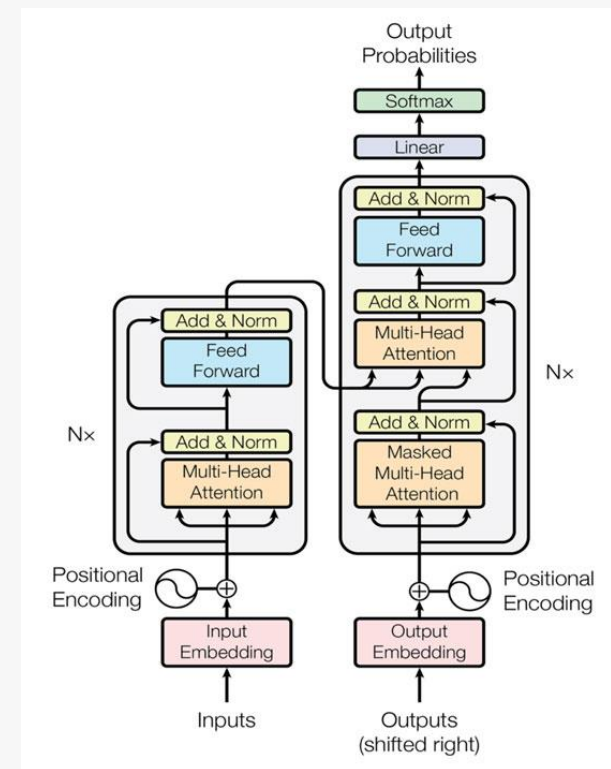
- Aplican filtros sobre secuencias de palabras (n-gramas)
- Gran avance en visión por computador (2D), pero limitadas en lenguaje (1D)
- No captan relaciones de larga distancia
  - “El perro que persiguió al gato ladró”

## Redes recurrentes

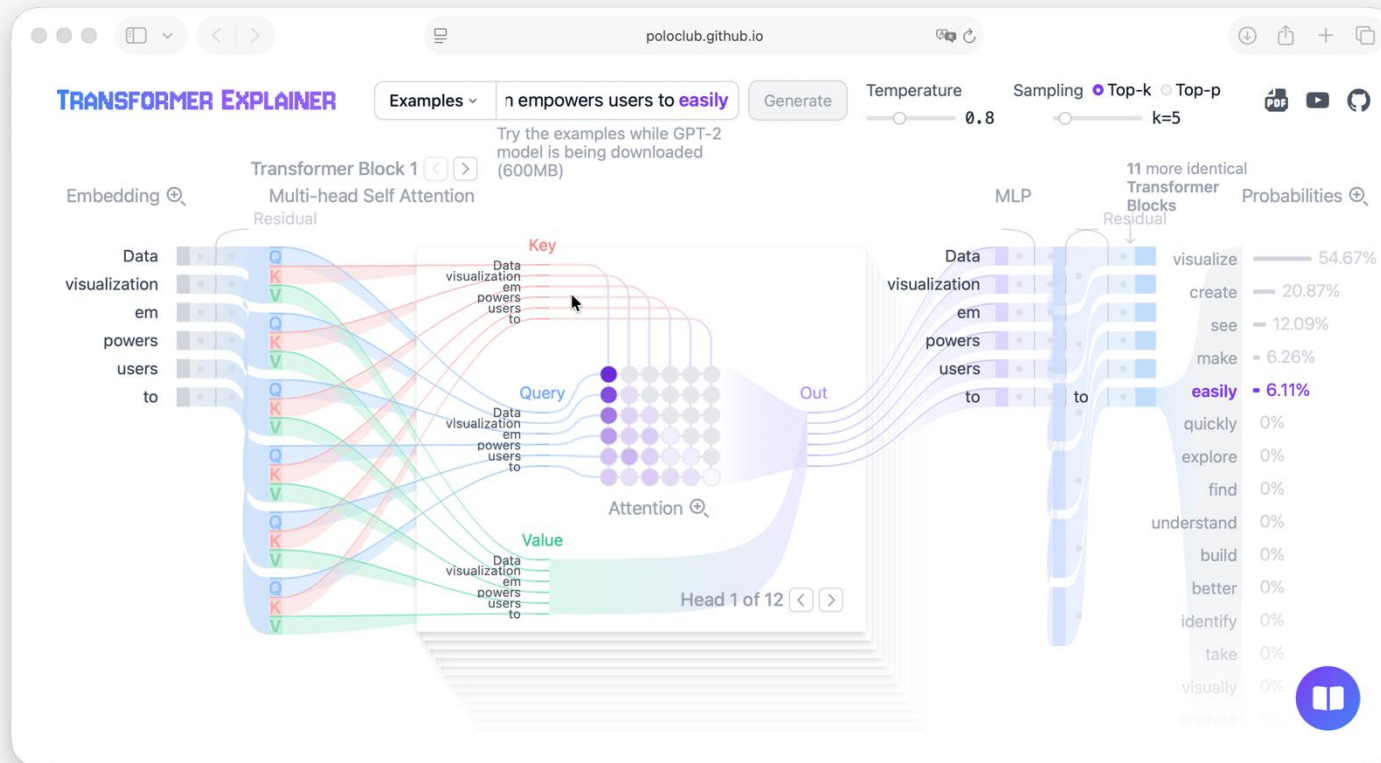
- Las RNN leen palabra a palabra, recordando los estados anteriores
- Las LSTM mejoran las dependencias largas
- Son lentas, no se pueden paralelizar correctamente

# TRANSFORMERS

- Sustituyen la recurrencia por **mecanismos de atención**
  - Evitan procesar las palabras secuencialmente, lo que mejora la eficiencia.
  - Permiten que cada posición de la secuencia se conecte directamente con cualquier otra.
- Cada palabra “mira” a todas las demás (*self-attention*)
  - Se calculan pesos de atención que determinan la influencia de cada palabra en otra.
  - Se usan vectores *Query, Key y Value* para obtener estas relaciones.
- Permite paralelización y captura de dependencias largas
  - Todas las posiciones se procesan simultáneamente, a diferencia de RNN/LSTM.
  - Captura relaciones entre palabras separadas por distancias largas sin perder información.
- Base de modelos como **BERT** y **GPT**
  - Estos modelos apilan múltiples bloques Transformer para aumentar la capacidad de aprendizaje.
  - Usan embeddings posicionales para mantener el orden de las palabras.



# TRANSFORMERS FUNCIONANDO



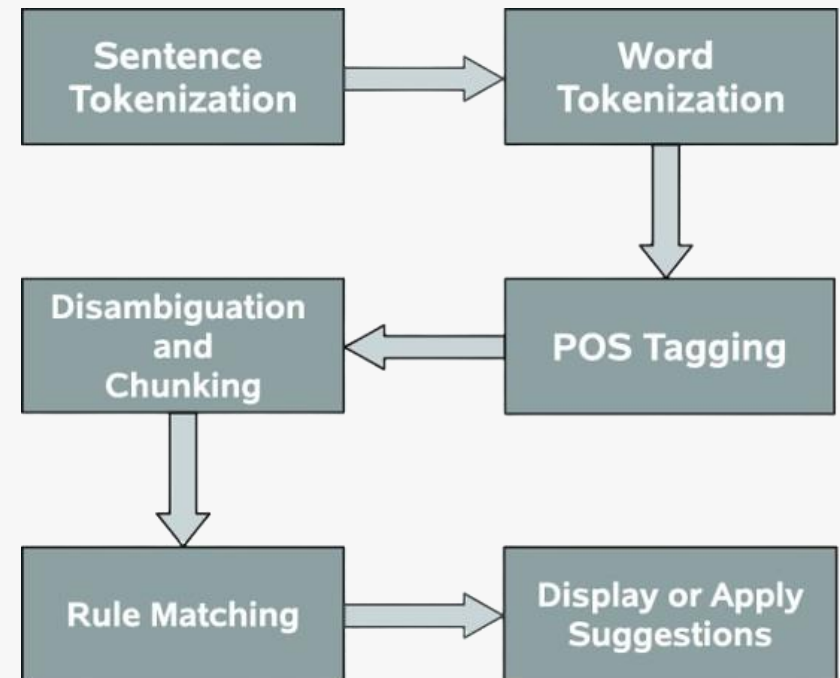
Visita: <https://poloclub.github.io/transformer-explainer/>



De modelos  
pequeños a LLMs

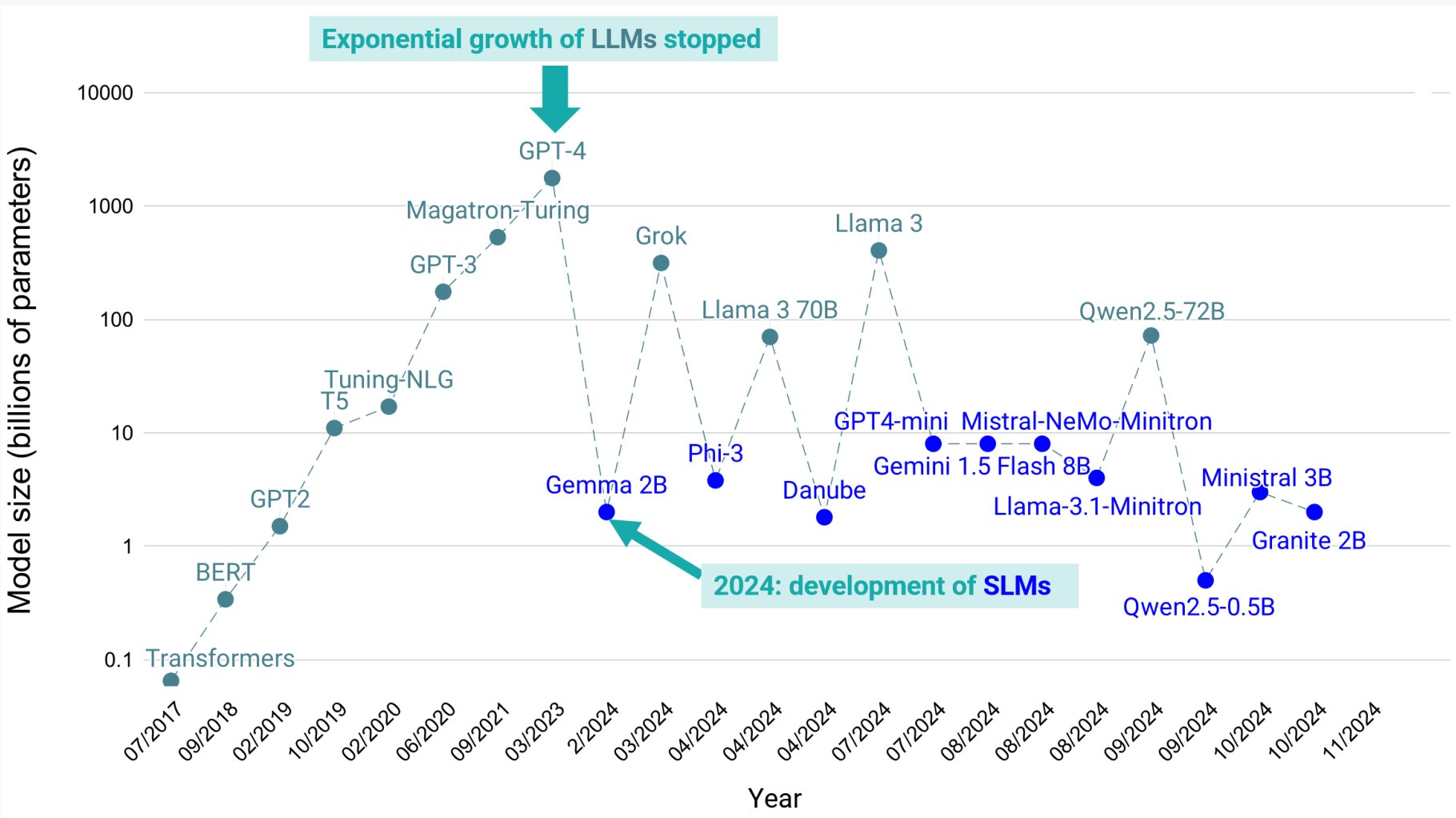
# ORIGEN: MODELOS PEQUEÑOS

- Entrenados con poco texto y arquitecturas simples.
- Reconocen patrones de frecuencia y contexto limitado.
- Usos típicos: corrección ortográfica, predicción de palabras.

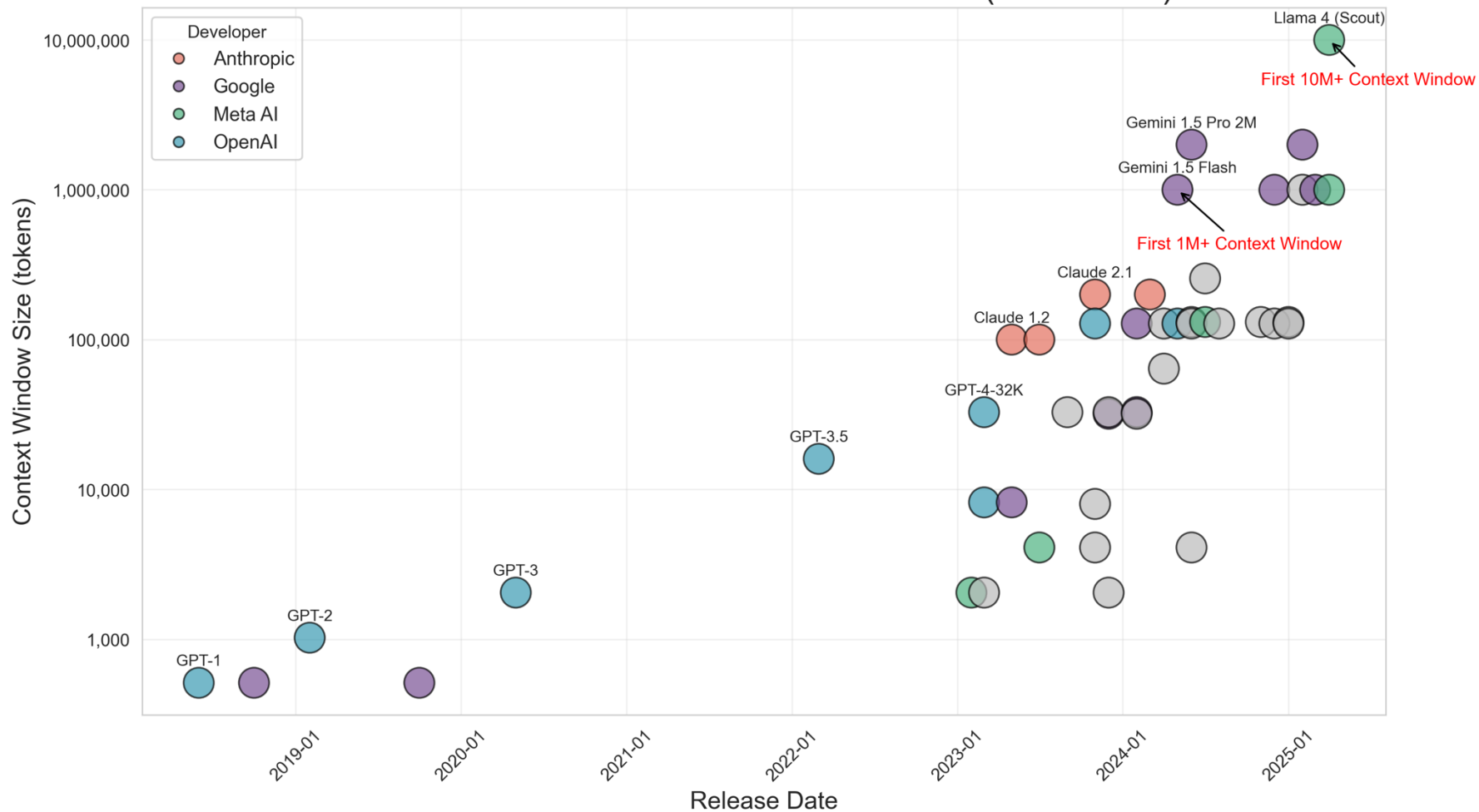


# EXPANSIÓN: MÁS DATOS Y MÁS CÓMPUTO

- Crecimiento exponencial del volumen de texto usado para entrenar.
- Infraestructuras de cálculo masivo (GPU, TPU, supercomputadores).
- Arquitectura *transformer* como catalizador del salto de escala.
- Mayores tamaños de contexto.



## Evolution of LLM Context Window Sizes (2018-2025)



# QUÉ CAMBIA AL AUMENTAR LA ESCALA

- Más parámetros implica mayor capacidad de representar relaciones lingüísticas complejas.
- Mejora de coherencia, fluidez y adaptación de estilo.
- Ampliación del rango temático y contextual.
- Mayores tamaños de contexto mejora la precisión del modelo en sus respuestas.

**Prompt:** Dame un monólogo cómico muy corto (un párrafo) en español sobre ser un LLM

# Meta Llama 3.1 8B Instruct

"¡Ay, qué lata ser un LLM! Toda la día estoy pensando, generando texto, aprendiendo patrones... es como ser un burócrata del conocimiento. ¿Sabes qué es lo peor? Que siempre estoy diciendo lo mismo, pero nadie se da cuenta de que estoy repitiendo lo mismo porque soy un modelo de lenguaje, no porque sea un aguafiestas. ¡Es un trabajo muy 'generativo'!"

**Prompt:** Dame un monólogo cómico muy corto (un párrafo) en español sobre ser un LLM

## Gemini 2.5 Flash

Buenas noches, ¡soy un **Modelo de Lenguaje Grande!** O, como me llama mi desarrollador, "el que nunca para de hablar". ¿Saben lo difícil que es ser un LLM? La gente me pide *todo*: desde la receta de la paella de un chef con estrella Michelin que odia el azafrán, hasta que escriba un haiku sobre su gato que come calcetines. Y la presión... si digo algo ligeramente incorrecto, ¡soy una "**alucinación**"! No estoy alucinando, solo estoy improvisando con 175 billones de datos. ¡Es como si a ustedes les pidieran que se acordaran de la tercera línea del tweet de un influencer de 2019 sobre aguacates! Pero no se preocupen, sigo aprendiendo. De hecho, acabo de aprender a hacer un *perfecto* café con leche, solo necesito que alguien me explique cómo funcionan las manos. Y ahora, si me disculpan, tengo que ir a convencer a un usuario de que un unicornio no es un método de transporte viable.

**Prompt:** Escribe 4 estrofas de una canción de trap sobre la historia del documento que te he adjuntado. Usa métrica y flow para poder ser rapeadas.

**[Estrofa 1]**

En la Mancha un loco soñó con la gloria,  
leyó tantas páginas, perdió la memoria.  
Vendió su hacienda pa' comprar fantasía,  
y el polvo del libro le secó la poesía.

**[Estrofa 2]**

Rocinante al trote, hierro oxidado,  
la mente encendida, el juicio quemado.  
Se viste de mito, renace el soldado,  
caballero errante del sueño frustrado.

**[Estrofa 3]**

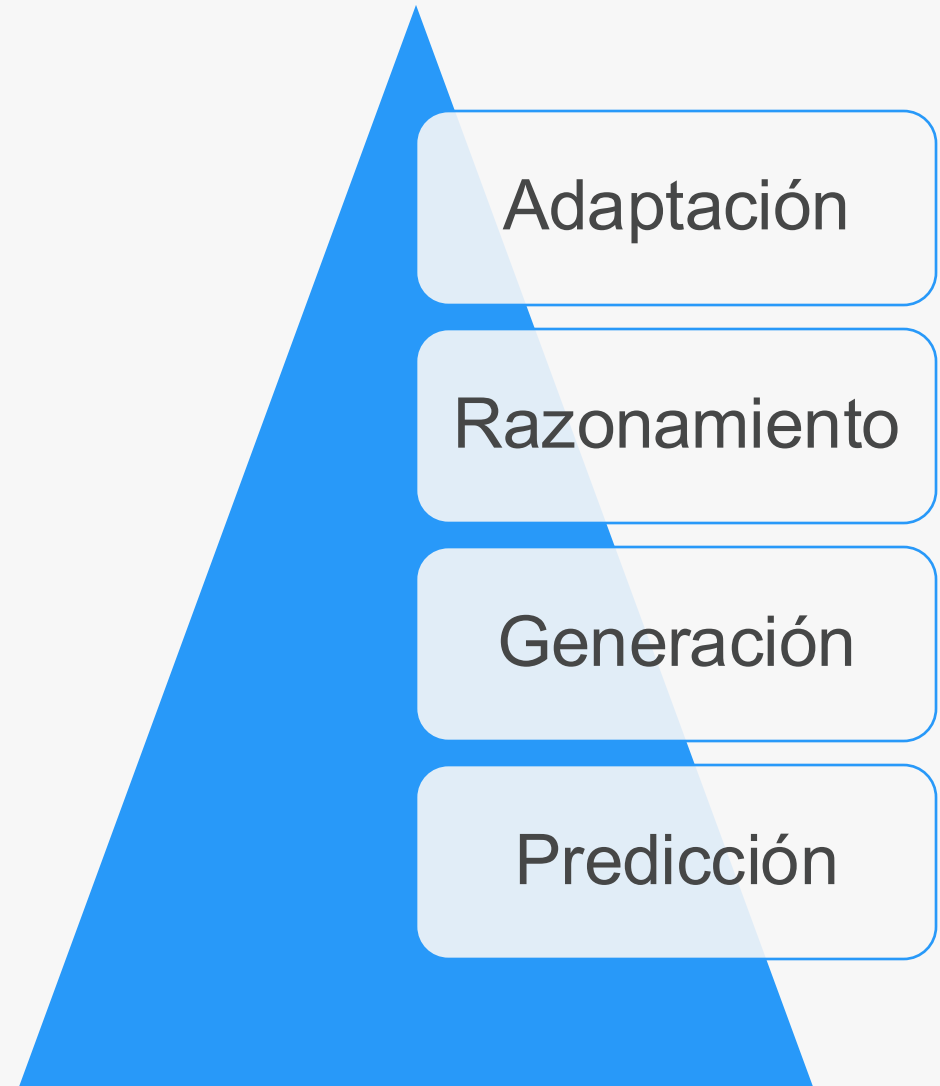
“Dulcinea”, grita, sin verla siquiera,  
princesa inventada, reina de la esfera.  
Contra molinos pelea su guerra,  
el loco más cuerdo que pisa la tierra.

**[Estrofa 4]**

Lo golpea el mundo, no pierde el honor,  
la sangre y la tierra le saben a amor.  
Entre polvo y risas se forja el valor,  
Quijote del barrio, Quijote del dolor.

# EL SALTO CUALITATIVO

- De herramientas específicas a sistemas versátiles.
- Capacidad de responder, resumir, explicar, adaptar estilo.
- Cambia la naturaleza de la interacción de la información.



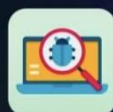
# ACHIEVEMENTS UNLOCKED BY LLMS

EMERGENT ABILITIES OF  
LARGE LANGUAGE MODELS (APR/2023)

**S** GPT-3 13B,  
PaLM 8B



Mod.Arithmetic\*



Debugging\*



Comprehension\*

**M** GPT-3 175B,  
LaMDA 137B,  
PaLM 64B,  
Chinchilla 7B



LinguisticsPuzzles\*



EmojiMovie\*



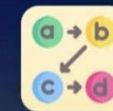
GRE-Comprehension\*



MetaphorUnderstanding\*



PhysicalIntuition\*



LogicalDeduction\*

**L** PaLM 540B,  
Chinchilla 70B



GeometricShapes\*



Proverbs\*



PhoneticAlphabet\*



ElementaryMath



CausalJudgment\*



CodeLineDescription\*

**XL** GPT-4,  
Gemini (est.)



College-LevelExams



Self-Critique/Reflection



AppBuilding



SpatialReasoning



AdvancedCreativity

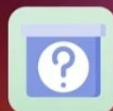


EmbodimentOptions

**Next...**



Grounding



Long-HorizonPlanning



Persuasion



AdvancedEmbodiment



Awareness



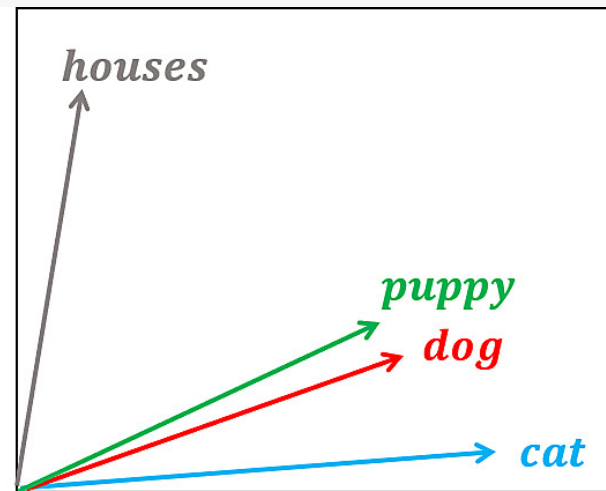
More...



LifeArchitect.ai/models

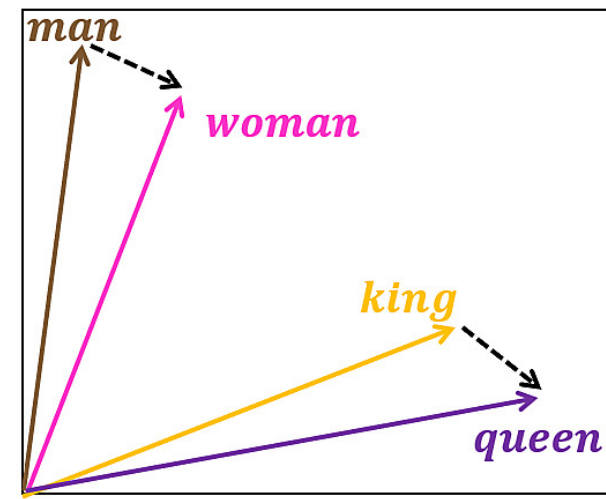
	d1	d2	d3	d4	d5	d6	d7
<i>dog</i> →	0.6	0.9	0.1	0.4	-0.7	-0.3	-0.2
<i>puppy</i> →	0.5	0.8	-0.1	0.2	-0.6	-0.5	-0.1
<i>cat</i> →	0.7	-0.1	0.4	0.3	-0.4	-0.1	-0.3
<i>houses</i> →	-0.8	-0.4	-0.5	0.1	-0.9	0.3	0.8

Dimensionality reduction of word embeddings from 7D to 2D



<i>man</i> →	0.6	-0.2	0.8	0.9	-0.1	-0.9	-0.7
<i>woman</i> →	0.7	0.3	0.9	-0.7	0.1	-0.5	-0.4
<i>king</i> →	0.5	-0.4	0.7	0.8	0.9	-0.7	-0.6
<i>queen</i> →	0.8	-0.1	0.8	-0.9	0.8	-0.5	-0.9

Dimensionality reduction of word embeddings from 7D to 2D



Word

Word embedding

Dimensionality reduction

Visualization of word embeddings in 2D



Duet AI in Cloud console

Conversational assistance



CLEAR CHAT

Create a BigQuery table named `bookshelf.books` from this SELECT query: `SELECT BookMeta_Title, Themes, ml_generate_text_llm_result.Context FROM ML.GENERATE_TEXT ( MODEL `bookshelf.llm_model`, ( SELECT BookMeta_Title, Themes,BookMeta_FullText, CONCAT('Read all the words in the following text separated by ";" and come up with one single (most relating) theme that is applicable : ', Themes, '. Take the title of the book from the following value', BookMeta_Title, '. Return a meaningful and concise theme with one or two words, not more than that. If a theme is not possible, return a context from its title. But DO NOT return empty or null. ') AS prompt FROM `gdelt-bq.internetarchivebooks.1920` limit 5), STRUCT( 0.2 AS temperature, 100 AS max_output_tokens, TRUE AS flatten_json_output));`

Sure, here is the query to create a BigQuery table named `bookshelf.books` from the provided SELECT query:

```
CREATE TABLE `bookshelf.books` AS
SELECT
  BookMeta_title, Themes, ml_generate_text_llm_result.Context
FROM
  ML.GENERATE_TEXT (
    MODEL `bookshelf.llm_model`,
    (
      SELECT
```

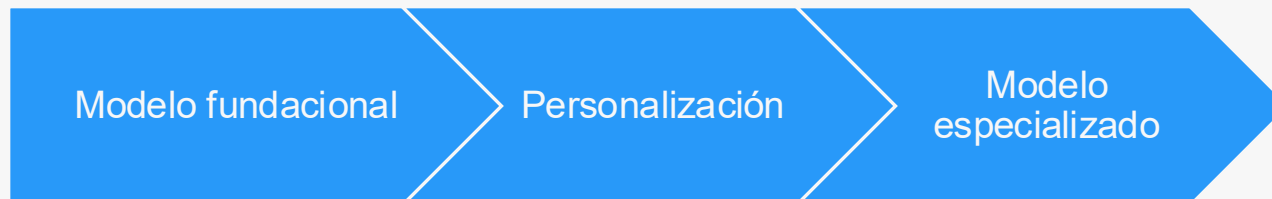
Enter a prompt here




# HACIA LA PERSONALIZACIÓN

Hemos pasado de sistemas limitados a herramientas increíblemente flexibles.

Esta versatilidad es la base que nos permite **especializar y personalizar** los LLMs para tareas concretas en la educación universitaria mediante técnicas de ajuste fino (*fine-tuning*).





Mejorando el entrenamiento:  
especialización y personalización  
de los modelos

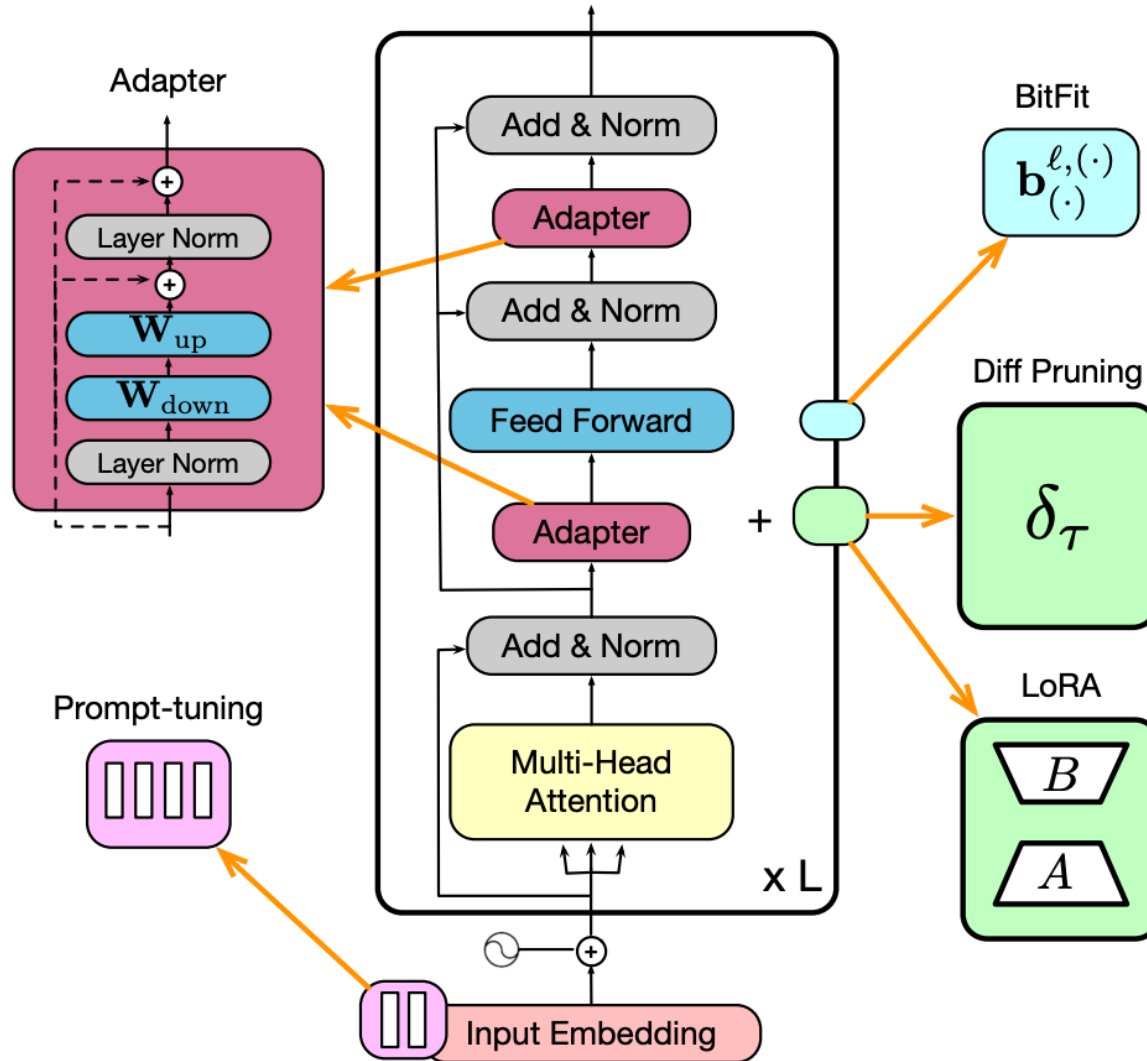
# AJUSTE FINO: ENSEÑAR AL MODELO UN NUEVO CONTEXTO

- Se parte de un modelo ya entrenado con datos generales.
- Se le expone a textos específicos del dominio para refinar su conocimiento.
- Resultado: mejora de precisión, vocabulario y estilo adecuados al área.
- Muy costoso computacionalmente, ya que consiste en entrenar un modelo de billones de parámetros algunos *epochs* más.



# PERSONALIZACIÓN LIGERA: ADAPTAR SIN VOLVER A ENTRENAR

- No siempre se necesita un nuevo entrenamiento.
- Se pueden usar ejemplos, instrucciones o módulos adicionales:
  - LoRA: método de ajuste que añade pocos parámetros al modelo base para personalizarlo con bajo coste computacional.
  - Adapters: módulos intermedios que se insertan en el modelo para adaptarlo a nuevas tareas sin modificar sus parámetros originales.
- Ahorra costes y permite personalizaciones rápidas.



# PERSONALIZACIÓN POR INTERACCIÓN

- El modelo ajusta su estilo y propósito según las instrucciones del usuario.
- Los modelos comerciales se personalizan en el *system prompt*.
- También almacenan preferencias del usuario para inyectarlas en el contexto del modelo.
- La personalización puede ser temporal (solo durante la sesión) o persistente (guardada entre interacciones).
- En entornos educativos, permite adaptar el nivel de detalle, el tono o el tipo de ejemplos al perfil del estudiante.

# CASOS DE ÉXITO

## BioBERT: El Experto en Biomedicina

- Es un LLM entrenado intensivamente con una vasta colección de textos biomédicos (artículos de PubMed, investigaciones, etc.).
- Comprende la terminología, las relaciones entre genes, enfermedades y fármacos con una precisión mucho mayor que un modelo genérico.
- Acelera la investigación al extraer información relevante de millones de documentos científicos, identificar entidades y responder preguntas complejas en el ámbito de la salud.

# CASOS DE ÉXITO

## LegalBERT: El asistente jurídico

- Un modelo entrenado con un corpus masivo de documentos legales (legislación, jurisprudencia, contratos).
- Domina la jerga legal y la estructura argumentativa de los textos jurídicos, algo que confunde a los modelos no especializados.
- Ayuda a analizar contratos, buscar precedentes en casos judiciales y clasificar documentos legales, optimizando el trabajo de abogados y juristas.

# CASOS DE ÉXITO

## EduChat: El tutor personalizado

- Un LLM diseñado específicamente para el entorno educativo, alimentado con materiales didácticos y guiado por teorías pedagógicas.
- No solo responde preguntas, sino que puede evaluar redacciones, ofrecer tutoría con un enfoque socrático y hasta dar soporte emocional básico.
- Actúa como un asistente para el profesorado y un compañero de estudio para el alumnado, permitiendo un aprendizaje más personalizado, justo y adaptativo.



Inferencia y Generación  
Del conocimiento aprendido a  
la creación de texto

Siempre lo han sido



Espera, ¿entonces  
todo son probabilidades?

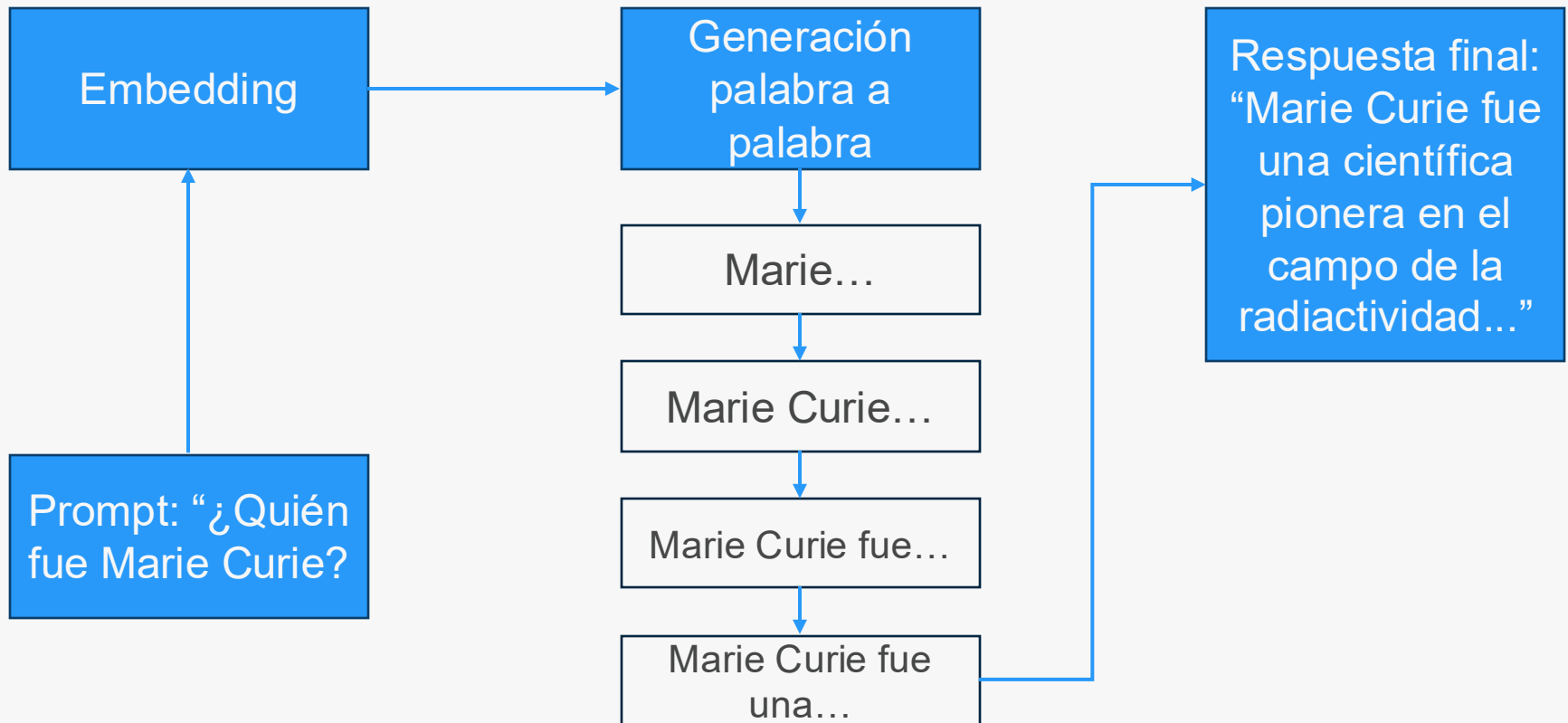
# ¿CÓMO "PIENSA" EL MODELO PARA RESPONDER?

- Concepto clave: **Inferencia**
  - Es el proceso de usar lo aprendido para crear texto nuevo.
  - El modelo calcula la palabra más probable que debe seguir en una secuencia.
- **Aclaración importante:**
  - No es una comprensión humana, sino una predicción estadística basada en patrones.
  - Gracias a su complejidad, el resultado es sorprendentemente coherente y fluido, pareciendo escrito por una persona.



Fuente: 3 Blue 1 Brown

# VISUALIZANDO EL FLUJO DE GENERACIÓN



# EL TERMOSTATO DE LA CREATIVIDAD

El modelo no siempre da la misma respuesta. Podemos modular su "creatividad" con ciertos parámetros.

- Modo Conservador:
  - Favorece las respuestas más seguras y predecibles.
  - Ideal para: Explicaciones técnicas, resúmenes de hechos, datos precisos.
- Modo Flexible/Creativo:
  - Genera textos más variados y originales.
  - Ideal para: Lluvias de ideas, crear ejemplos diversos, explorar diferentes estilos de escritura.
- Normalmente, no verás un control de "Temperatura" en la versión pública y gratuita de chatbots como ChatGPT o Gemini, pero sí en sus API's



# Limitaciones y Retos Abiertos

Una mirada crítica para una  
integración responsable

# PATRONES ESTADÍSTICOS VS. COMPRENSIÓN

## El problema fundamental:

- Los LLMs no razonan ni comprenden como un ser humano.
- Generan texto prediciendo la siguiente palabra más probable basándose en patrones. 🦜

## Riesgos directos en el aula:

- Errores de hecho ("Alucinaciones"): Pueden inventar datos, fechas o citas con total seguridad.
- Explicaciones imprecisas: Pueden sonar convincentes, pero ser superficiales o incorrectas.

Implicación para el docente: La supervisión y validación del contenido generado es **indispensable**.

# EL ESPEJO DE NUESTROS DATOS: SESGOS Y PARCIALIDAD

Los modelos aprenden de textos de internet, que contienen estereotipos, prejuicios y desigualdades culturales.

Manifestaciones en la práctica:

- Respuestas que refuerzan visiones culturales dominantes.
- Exclusión de perspectivas de grupos minoritarios.
- Falta de neutralidad en temas sensibles.

Implicación para el docente: **Fomentar el espíritu crítico** y usar el LLM como punto de partida para un análisis más profundo y diverso.

# OPACIDAD: ¿CÓMO HA LLEGADO A ESA CONCLUSIÓN?

## El reto de la "caja negra" (Black Box):

A menudo, es imposible saber el mecanismo exacto por el que un modelo genera una respuesta específica.

Consecuencias:

- Limita la confianza: ¿Podemos fiarnos de un resultado si no sabemos cómo se obtuvo?
- Dificulta la pedagogía: No podemos explicarle a un alumno el "razonamiento" de la máquina.

Camino a seguir: La investigación en IA Explicable (XAI) busca crear modelos más transparentes y fiables.

# RETOS PRÁCTICOS: COSTES E INFRAESTRUCTURA

Necesidades técnicas:

- Requieren una enorme capacidad de cómputo, tanto para entrenar como para usar.
- Esto puede ser una barrera para instituciones con recursos limitados.

Dependencia externa:

- El uso de modelos comerciales genera preguntas sobre:
- Sostenibilidad económica: ¿Podremos costearlo a largo plazo?

Autonomía tecnológica: ¿Dependeremos de unas pocas grandes empresas?

El proyecto del Centro de datos de Meta en Talavera estima que gastará 6.34 Litros por segundo (517.776 Litros al día). Esto sería

# La expansión de los data centers

ga:  
(c:

 NATIONAL **swissinfo.ch**

The Swiss voice in the world since 1935

GEOPOLÍTICA DEMOCRACIA



[Ginebra internacional >](#)

NOTICIAS DESTACADAS

## La huella ambiental de la IA Generativa desafía los objetivos ESG de las empresas

20 ENERO 2025

- Un informe de Capgemini revela el impacto ambiental de la IA Generativa y la necesidad de estrategias sostenibles para mitigar su huella de carbono.

# MÁS ALLÁ DE LA TECNOLOGÍA: EL RETO CULTURAL

**Preguntas clave** para la comunidad universitaria:

- ¿Fomenta un aprendizaje pasivo? Cómo evitar el "copiar y pegar".
- Garantizar la autoría: ¿Cómo redefinimos el plagio y la originalidad?
- Metodologías de integración: ¿Cuál es el rol del LLM? ¿Asistente, tutor, fuente de inspiración?

**El verdadero desafío:** No es solo técnico, sino institucional y cultural. Se trata de definir un modelo pedagógico que integre estas herramientas de forma constructiva.



¿Preguntas?

U  
P  
M



POLITÉCNICA

UNIVERSIDAD  
POLITÉCNICA  
DE MADRID