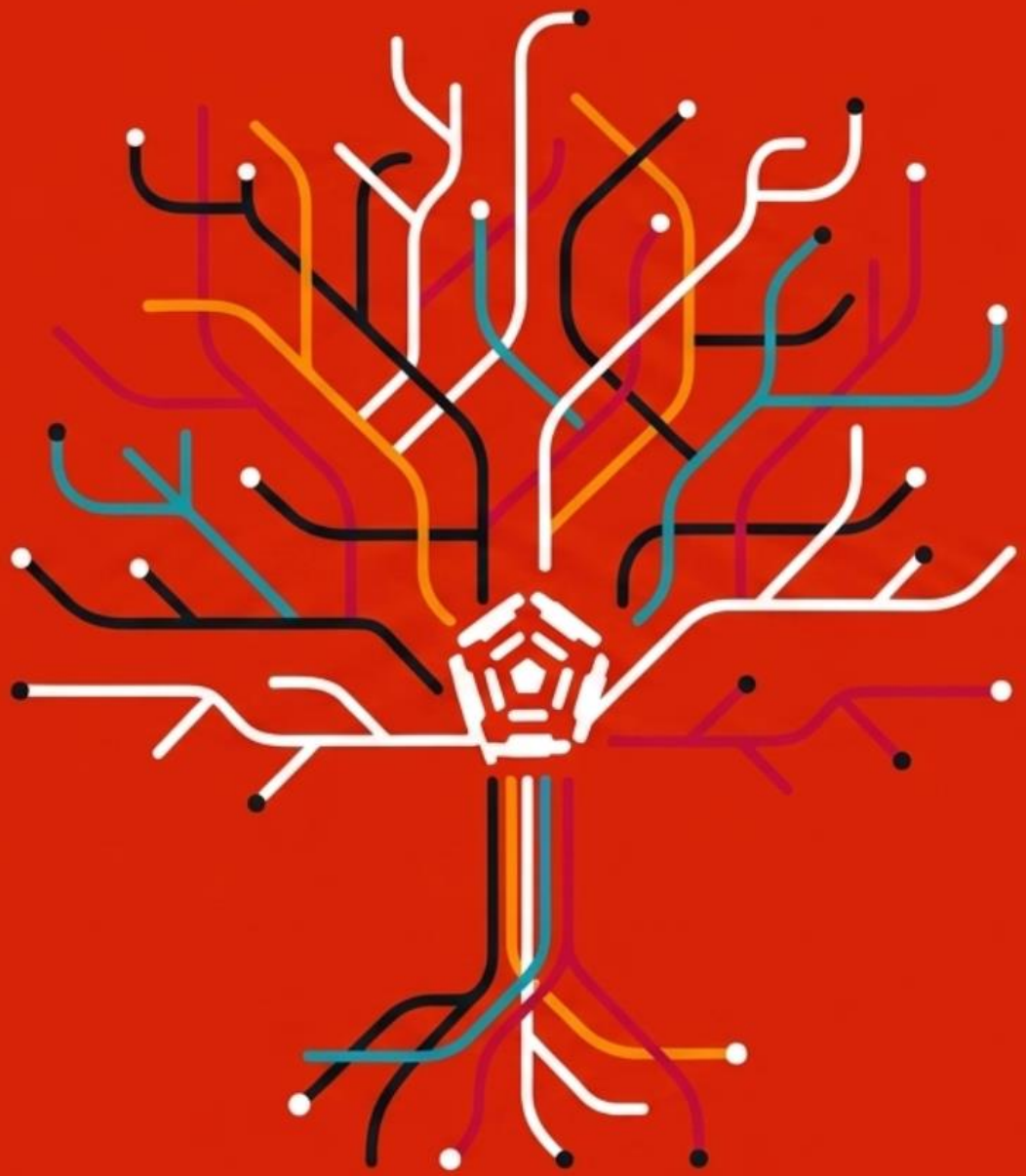
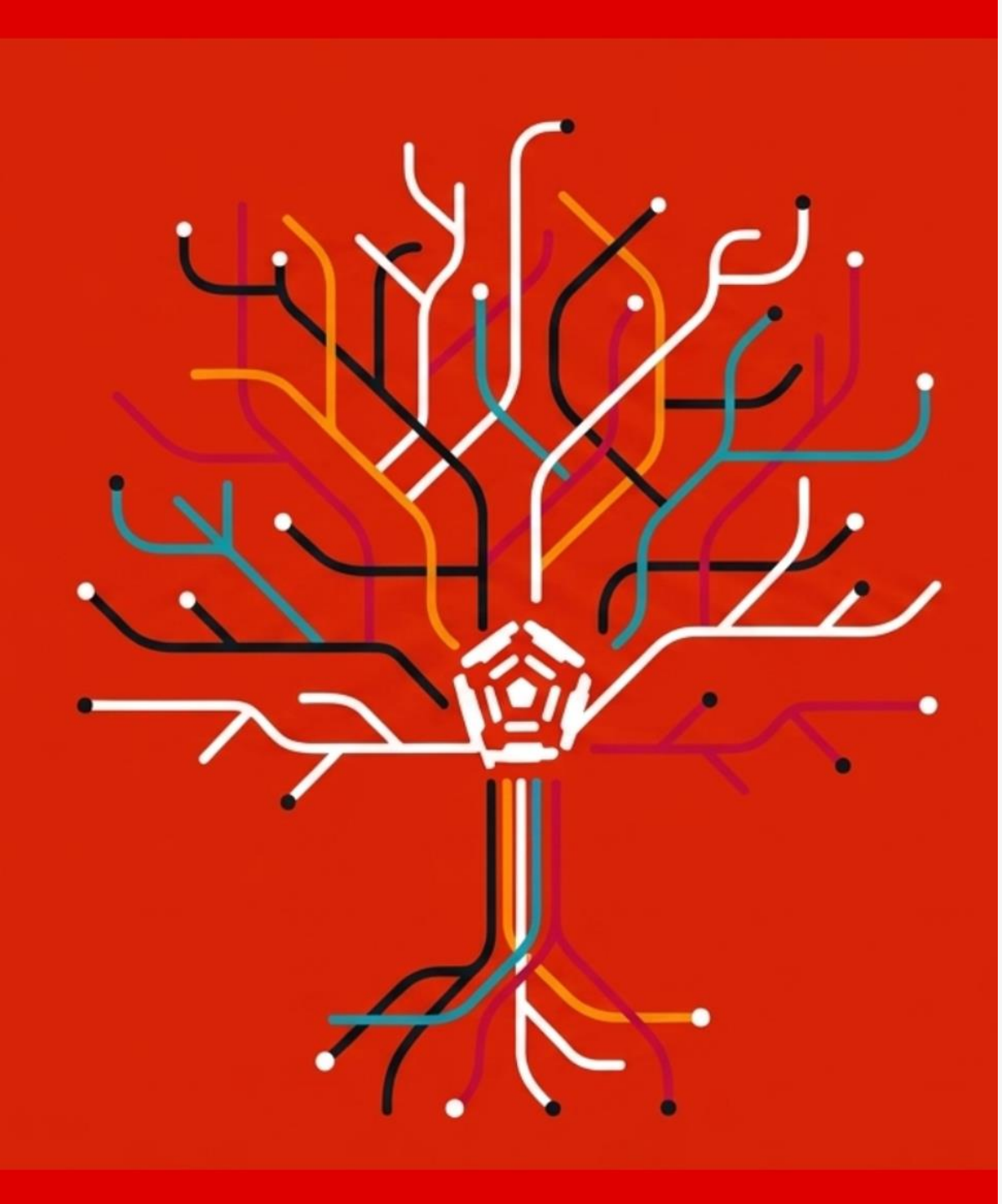


Universidad de Zaragoza | Curso 2026

# ¿QUÉ TIENE LA INTELIGENCIA ARTIFICIAL PARA MÍ ARCHIVO?

## De la teoría a la práctica.



**Cursos Extraordinarios**
**Universidad de Zaragoza 2026**

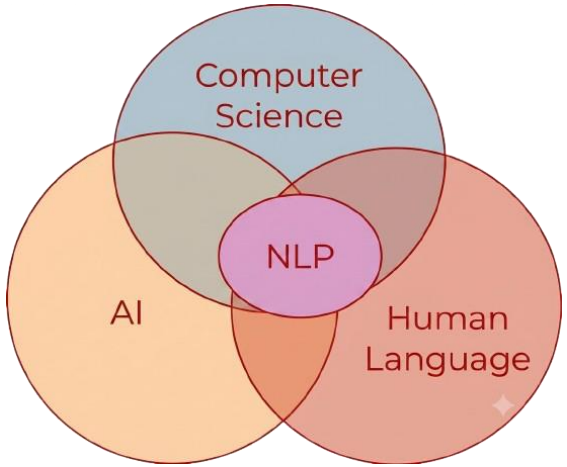
¿Qué tiene la IA para mi archivo?
De la teoría a la práctica

# Introducción al Procesado del Lenguaje Natural

# Introducción al Procesado del Lenguaje Natural

- **¿Qué es PLN?**
  Tecnologías que permiten a las máquinas procesar el lenguaje humano: escrito, oral o visual

- **Relevancia:**
  - Componente fundamental de los sistemas de IA




Capacidades de un sistema de IA
- ❑ Percepción
- ❑ Aprendizaje
- ❑ Representación del conocimiento
- ❑ Razonamiento

# Introducción al Procesado del Lenguaje Natural

- **Comprensión**

  El objetivo final es comprender el mensaje codificado en el lenguaje
  - Saber qué se dice, qué se hace o qué sucede.
  - Descubrir el sentido profundo de algo.
  - Entender conceptos y procesos para explicarlos o describirlos.

Proporciona herramientas para representar el *conocimiento*

**REPRESENTACIONES SEMÁNTICAS**

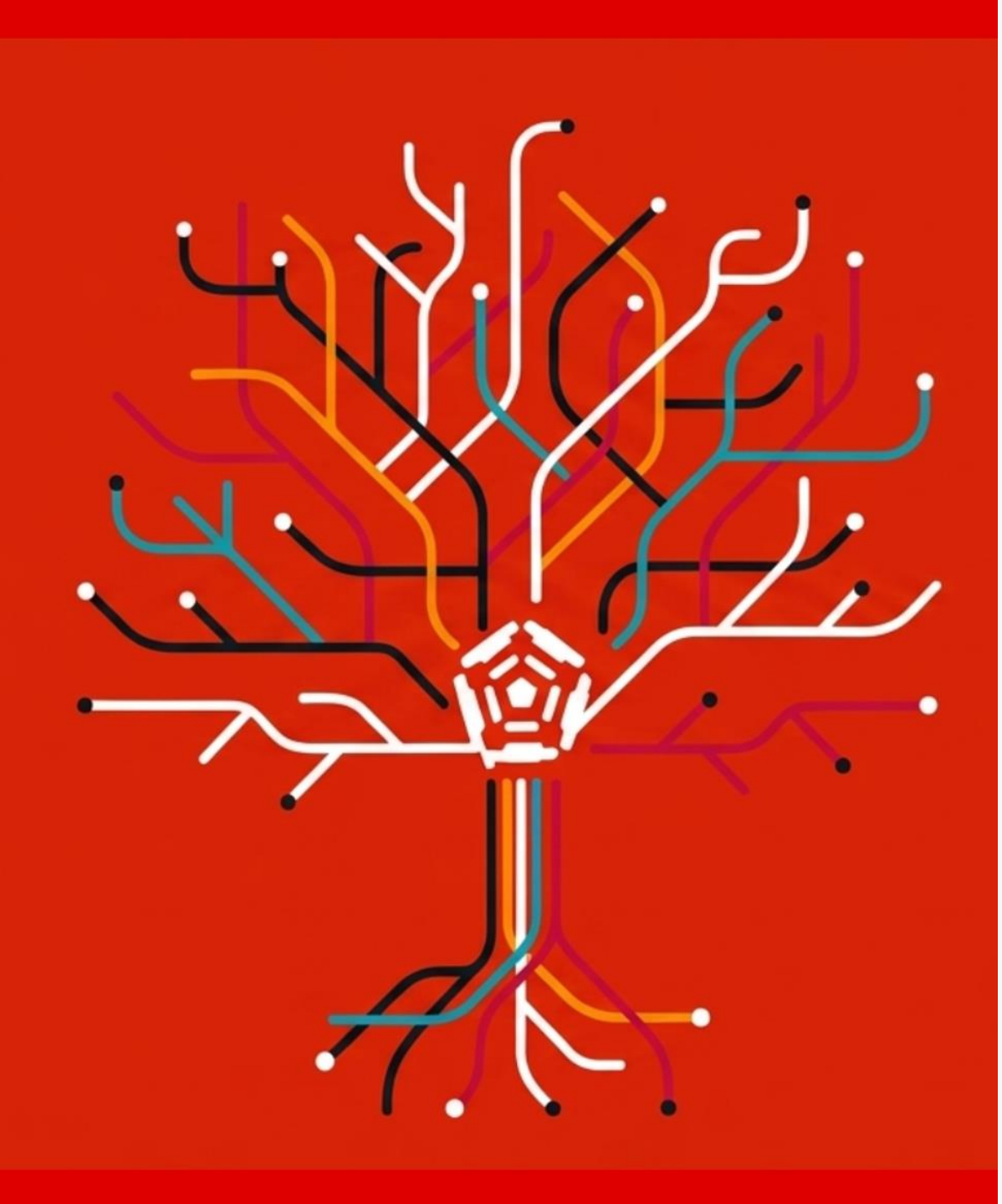
# Introducción al Procesado del Lenguaje Natural

- **Objetivos:**

✓ Representar el significado de las palabras
✓ Definir una medida de similitud semántica entre palabras
✓ Una representación numérica densa: “embeddings”

Obtener un espacio de representación compacto donde la posición de las unidades contenga información semántica:

***Espacio semántico: Modelos de Lenguaje***



**Cursos Extraordinarios**
**Universidad de Zaragoza 2026**

¿Qué tiene la IA para mi archivo?
De la teoría a la práctica

# Modelos de Lenguaje

# ¿Qué es un modelo de lenguaje (LM)?

- Representación matemática / computacional de la utilización de la lengua
- Usados para comprender, generar y predecir texto en función de contexto
- Enfoque probabilístico:

  ¿Qué palabra es más probable tras …

  *"documento digitalizado en el"*

  - ❑ "horno"
  - ❑ "archivo"

# Tipos de modelos de lenguaje

## Basados en Reglas:

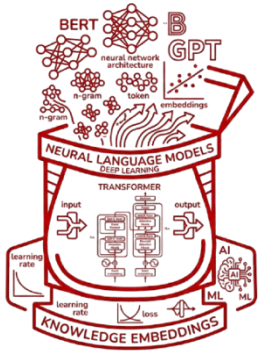
Gramáticas manuales creadas por expertos.
Rígidos y estáticos.

## Basados en Datos:

Aprenden patrones a partir de grandes corpus.
Flexibles y escalables.

# Evolución de los modelos de lenguaje






## Bag-of-Words

Cuenta palabras ignorando el orden.

Búsquedas básicas de palabras clave, clasificación de textos, análisis de sentimiento.

## Estadísticos

Probabilidades de ocurrencia de palabras o secuencias de palabras.

Transcripción de voz, traducción automática y generación de texto.

## Neuronales

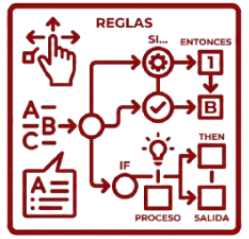
Patrones y representaciones del lenguaje.

Representaciones semánticas en espacios vectoriales.

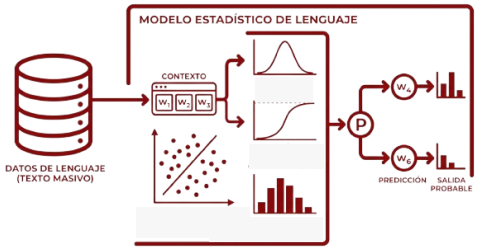
Útil para casi todo ...

# Evolución de los modelos de lenguaje

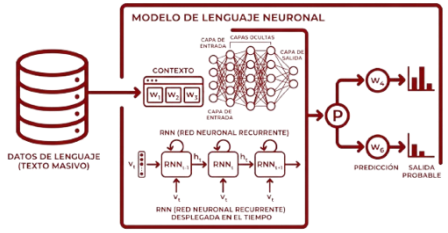
**1950**
- Basados en Reglas



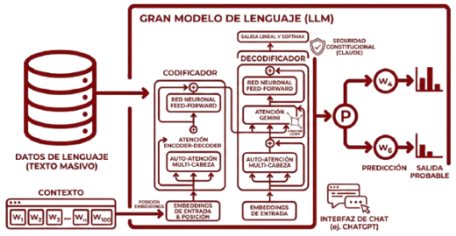

**1980**
- Estadísticos




**2000**
- Neuronales



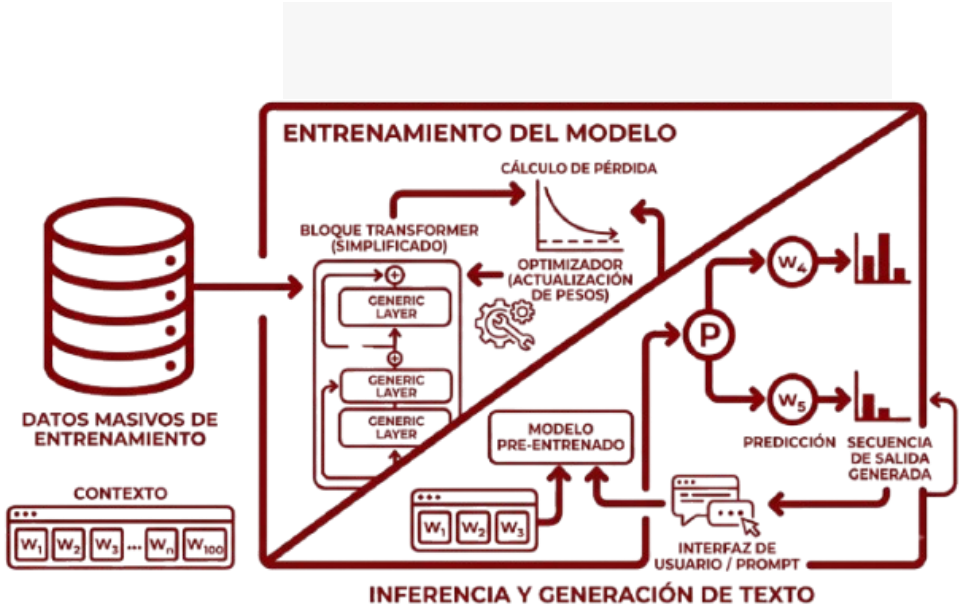

**2020**
- Large Language Models (LLM)



# Creación y uso de los modelos de lenguaje

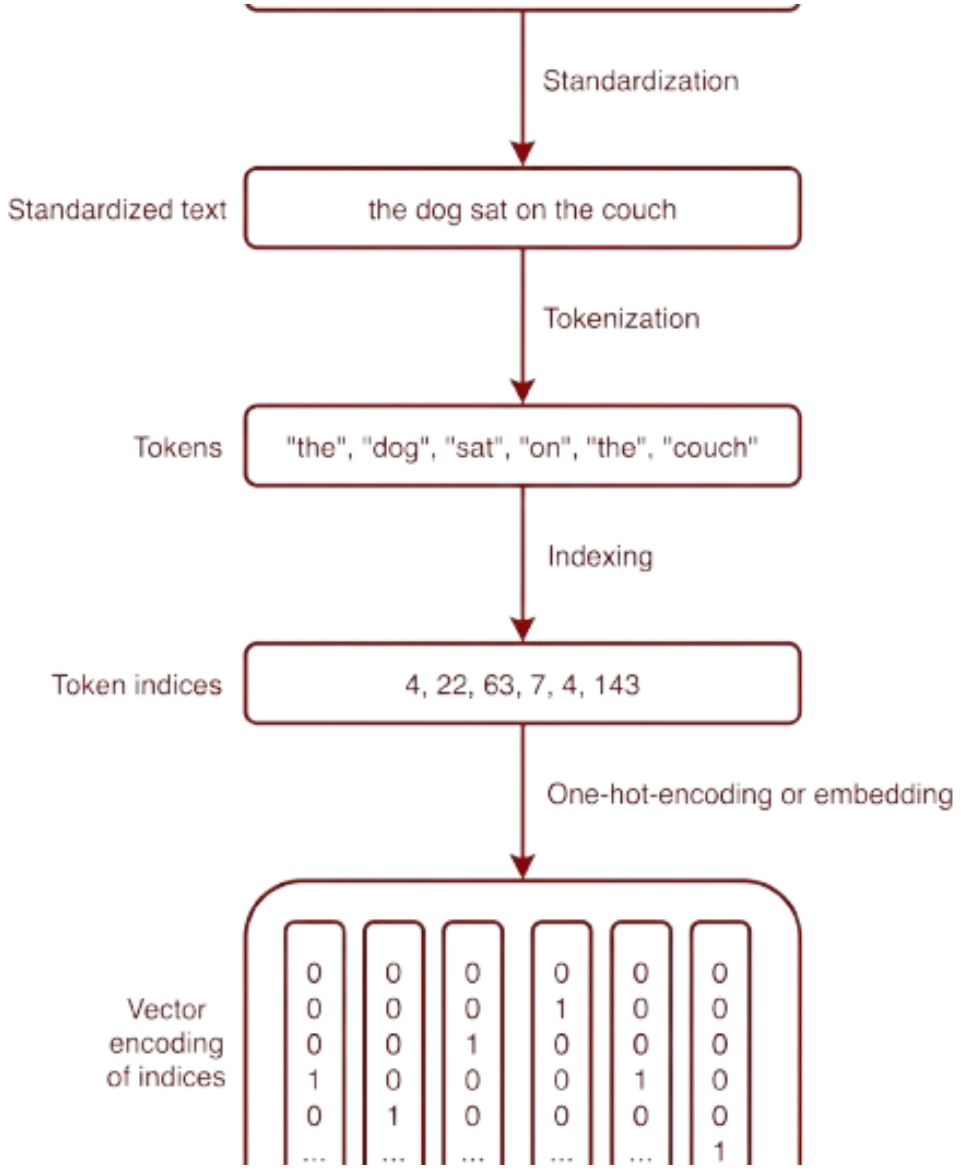
- **Creación (Entrenamiento):**
  - Se procesan millones de documentos (Corpus) para que el modelo encuentre patrones estadísticos.
  - Es imprescindible tratar (“curar”) los textos: Limpieza, preprocesado y representación.

- **Uso (Inferencia):**
  - El modelo "congelado" recibe una consulta y genera una respuesta basada en lo aprendido.
    - Probabilidad de una secuencia de texto
    - Generar la secuencia de palabras más probable
    - …



# Representación del Texto en los ML

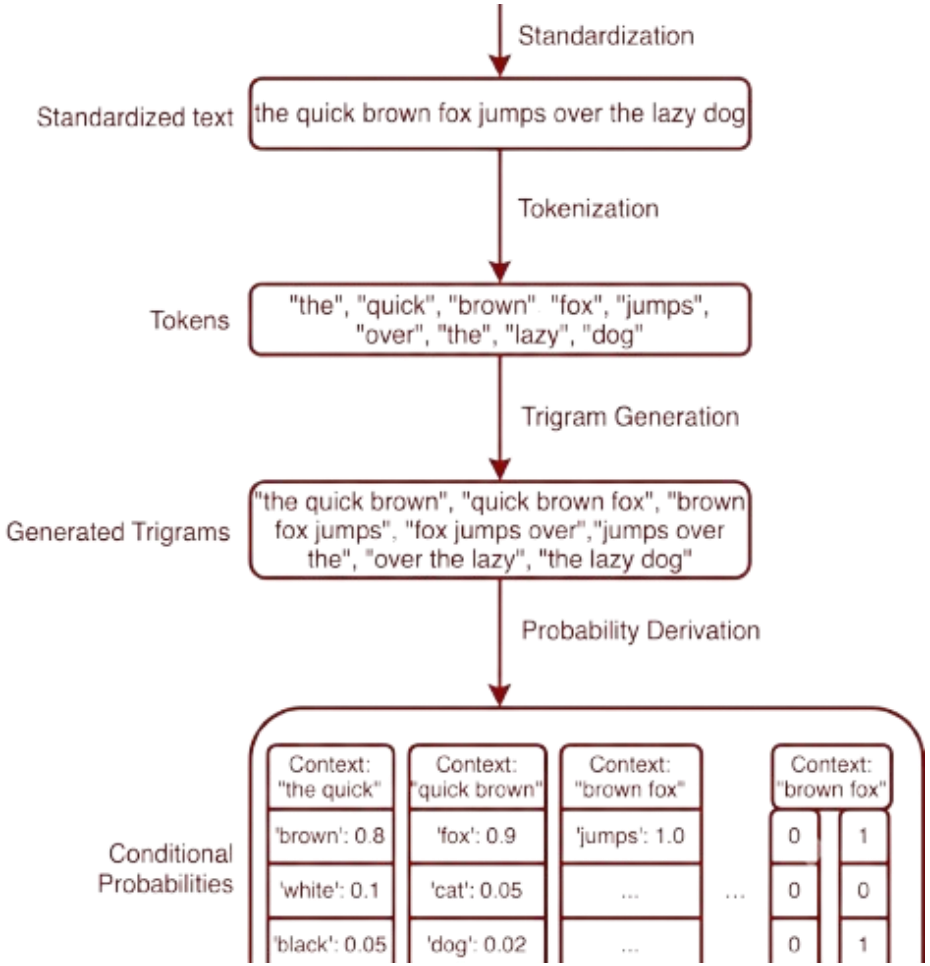
- **Tokens**:
  - Las máquinas no entienden letras, procesan números.
  - El texto debe representarse con secuencias de números.
  - Unidades de representación:
    - Palabras
    - Sílabas
    - Caracteres
    - Subpalabras
      - BPE (Byte Pair Econding)
        - GPT-2
      - WordPiece
        - BERT



# Modelos de Lenguaje Estadísticos

- **N-gramas:**
  - Basados en estadísticas (frecuencias relativas de aparición) de secuencias de palabras.
  - Estimadas en grandes repositorios textuales.
  - N define la profundidad de las estadísticas:
    - Unigramas: Frecuencia de aparición de cada unidad (“Expediente”)
    - Bigrama: Frecuencia de aparición de cada pareja de palabras (“Expediente sancionador”)
    - Trigrama: Combinaciones de tres palabras (”Expediente sancionador administrativo”)
    - …
  - Codifican la probabilidad de aparición de una palabra, dadas las N-1 anteriores

    Prob (“administrativo” | “expediente sancionador”)



# Modelos de Lenguaje Estadísticos

- **Generación:**
  - En esencia un modelo de lenguaje moderno es un sistema aleatorio.
  - La generación de los textos no es determinista
  - Existen ciertos parámetros que controlan el grado de aleatoriedad del sistema

**Temperatura**

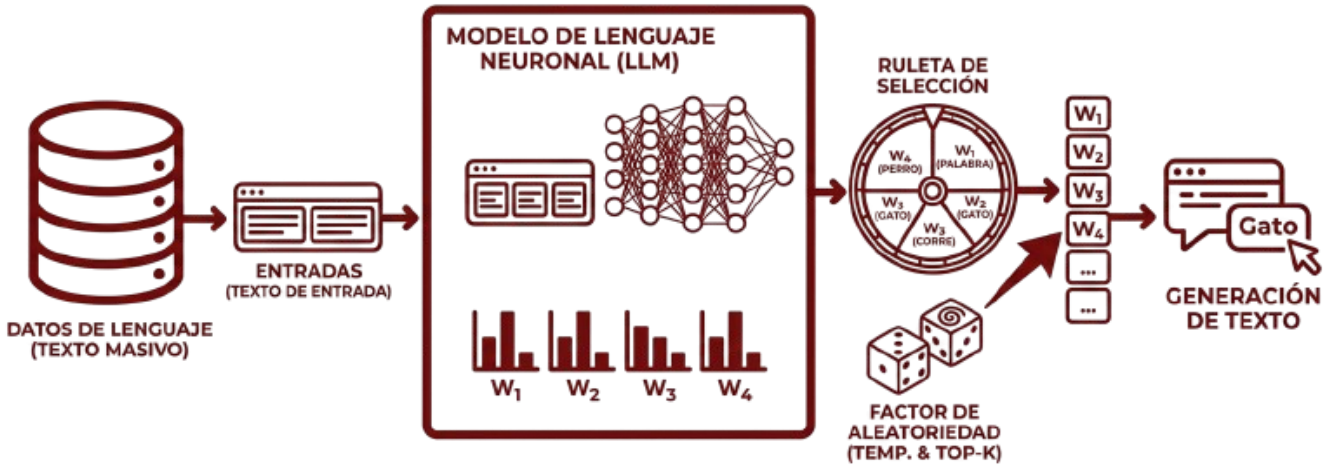
Bajo = Determinista
Alto = Creativo

**Top-k**

El modelo elige entre las k palabras más probable

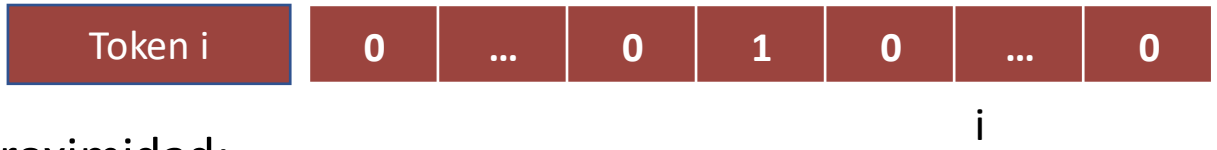
**Top-p**

Selecciona el conjunto de palabras que acumulen una probabilidad p



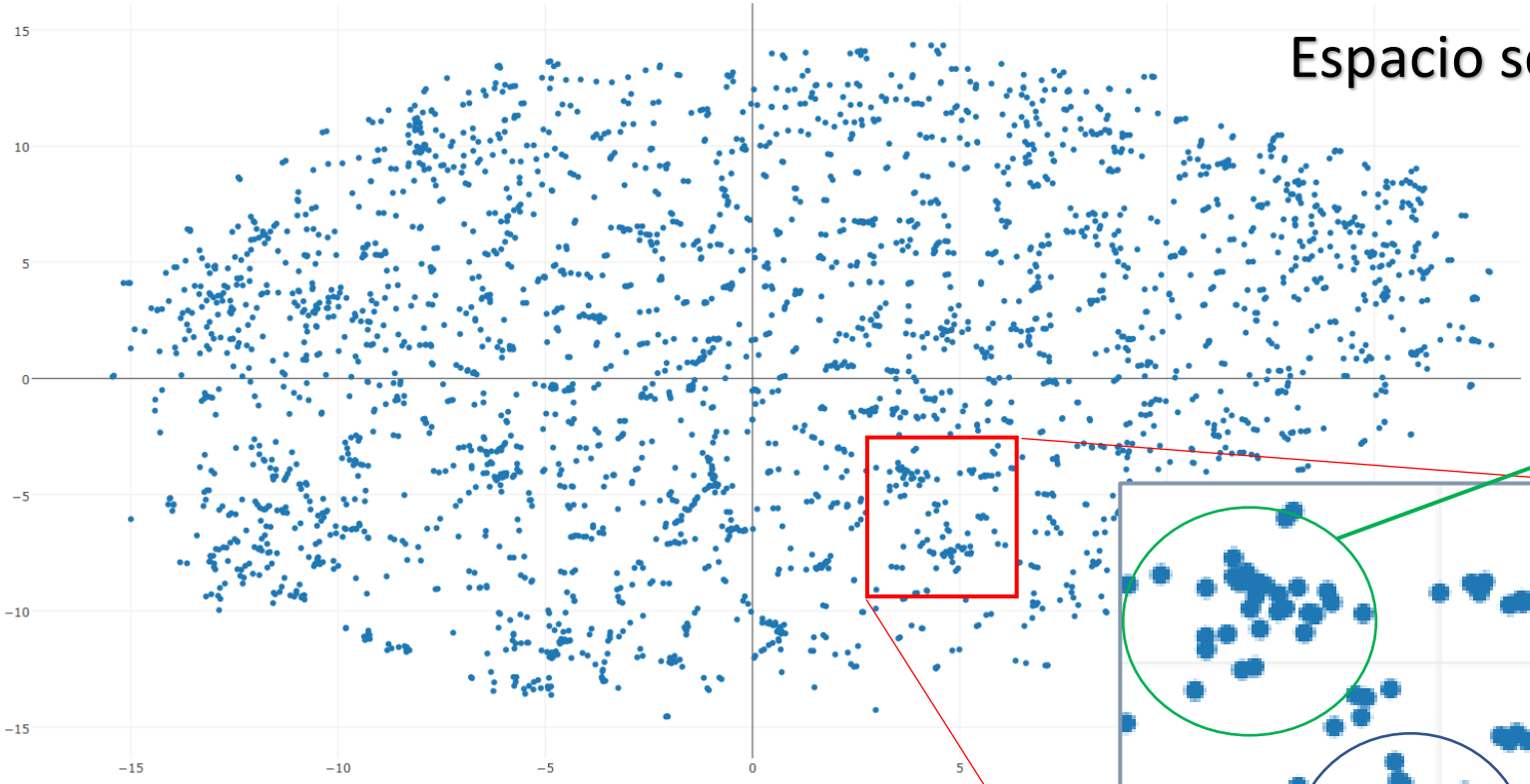
# Modelos de Lenguaje Neuronales

- **Representación de la Información:**
  - Representaciones Vectoriales:
    - Las palabras/tokens se convierten en vectores (coordenadas en un espacio multidimensional).
    - **One-hot encoding**
      - Una posición a 1 y el resto a 0
        - Ausencia de similitud por proximidad:
          - Todos están a la misma distancia de todos, no hay vectores más próximos a otros.
        - Muy alta dimensionalidad:
          - Precisa de tantas dimensiones como posibles tokens (decenas o centenares de miles)




    - **Embeddings**
      - Busca mantener **relaciones semánticas**:
        - Palabras con significados similares (ej. "Manuscrito" y "Códice") se ubican matemáticamente cerca.
        - Permite obtener similitud por proximidad

# Espacios de Representación Semánticos




Espacio semántico: *embeddings*

Sótano
Pasillo
Patio
Ático
Salón
Balcón
…

Fútbol
Béisbol
Rugby
Baloncesto
Tenis
Balonmano
Esquí
….

La proximidad semántica entre unidades lingüísticas se representa por la proximidad en el espacio

# Espacios de Representación Semánticos

- **<u>Principio de Funcionamiento:</u>**
  - Hipótesis Distribucional:

**PALABRAS SIMILARES OCURREN EN CONTEXTOS SIMILARES**

- Se basa en descubrir patrones estadísticos de las palabras a partir de los cuales descubrir diferencias o similitudes entre ellas.

- El significado de una palabra se obtiene de la distribución de palabras que están alrededor de ella.

- **<u>Si dos palabras tienen distribuciones estadísticas similares (en función del contexto), tienen significados similares.</u>**

  - ”Códice” se situará próxima a “Manuscrito”
  - Por el contexto, desambiguará y sabrá que “Audiencia” puede referirse a “tribunal” o a ”espectadores”

# Espacios de Representación Semánticos

- **Construcción de Espacios Semánticos:**

John Rupert Firth, "You shall know a word by the Company it keeps"
"Similar words occur in similar contexts"
Ludwig Wittgenstein, "The meaning of a word is its use in language"

Hay una botella de *Belikin* sobre la mesa
A todo el mundo le gusta la *Belikin*
No bebas *Belikin* si tienes que conducir
La *Belikin* se fabrica con granos de cebada germinada

¿qué podemos deducir sobre la palabra *Belikin*?

Miramos las palabras que acompañan
Buscamos la similitud semántica con otras palabras ya conocidas

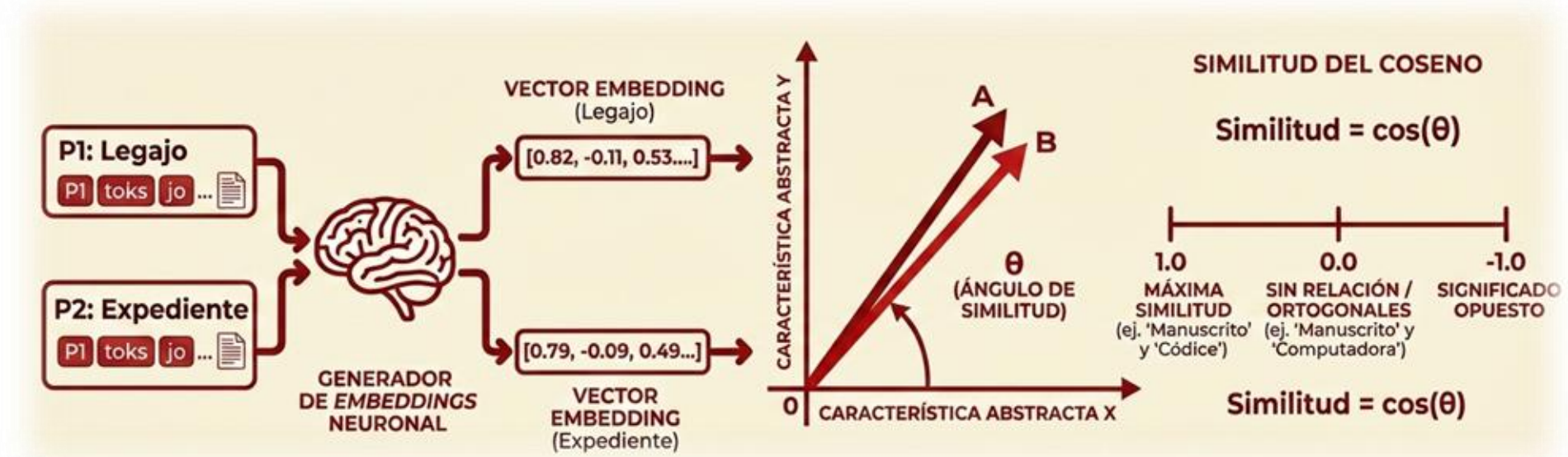
# Espacios de Representación Semánticos

- **Medida de Similitud Semántica:**
  - Cada palabra viene dada por su correspondiente *embedding*
    - Coordenadas en un mapa
      - Palabras con significados o contextos similares son "vecinas".
    - Medir la similitud semántica consiste en calcular la "distancia" o el "ángulo" entre embeddings (vectores).

  - **Métrica de Similitud Coseno:**
    - La distancia no se mide en línea recta (Euclídea), sino en función del ángulo entre vectores.
    - Idea Intuitiva:
      - Un documento de 100 páginas habla sobre "Juicios de Residencia" y un acta de 2 páginas habla de lo mismo
        - Sus vectores tendrán longitudes muy diferentes (distancia euclídea grande), pero apuntarán en la misma dirección (ángulo casi cero).
      - Esta métrica ignora el tamaño del texto y premia el contenido temático.

# Espacios de Representación Semánticos

- **Ejemplo Práctico de Similitud Semántica:**
  - Buscando en un conjunto de documentos “Legajo”, si la búsqueda no es semántica, aquellos documentos que contienen “Expediente” o “Manuscrito” no aparecerán.
  - Con el uso de embeddings semánticos y la similitud coseno:
    - `Similitud("Legajo", "Expediente") = 0.89` (Alta -> coincidencia)
    - `Similitud("Legajo", "Factura de la luz") = 0.08` (Baja -> descarte)



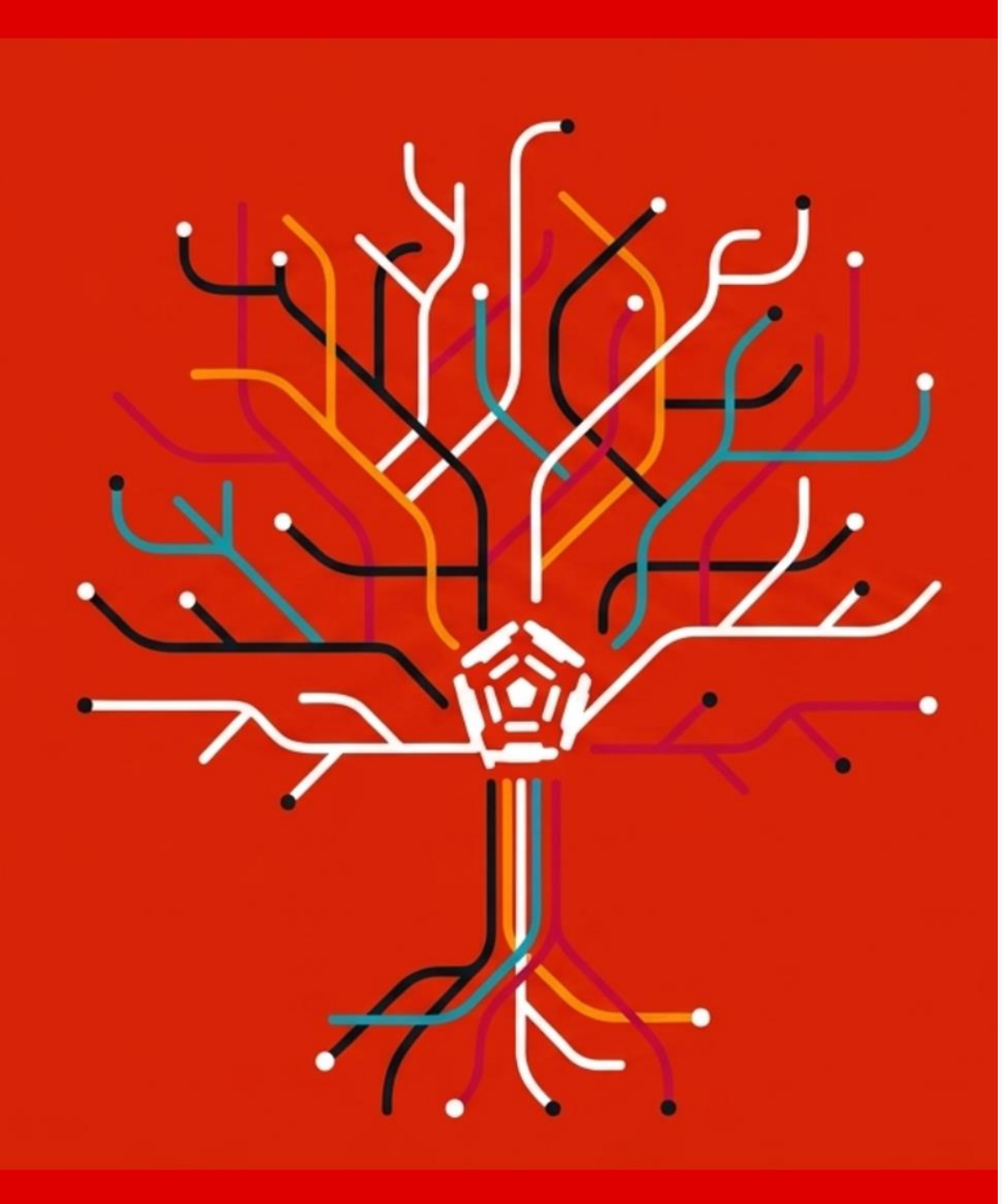
# Espacios de Representación Semánticos

- **Ejemplo Práctico de Similitud Semántica:**

  - **Utilidad Práctica:**

    - **Búsqueda conceptual**: Encontrar documentos que hablan de lo mismo, aunque usen palabras distintas (sinónimos históricos, evolución del lenguaje administrativo).

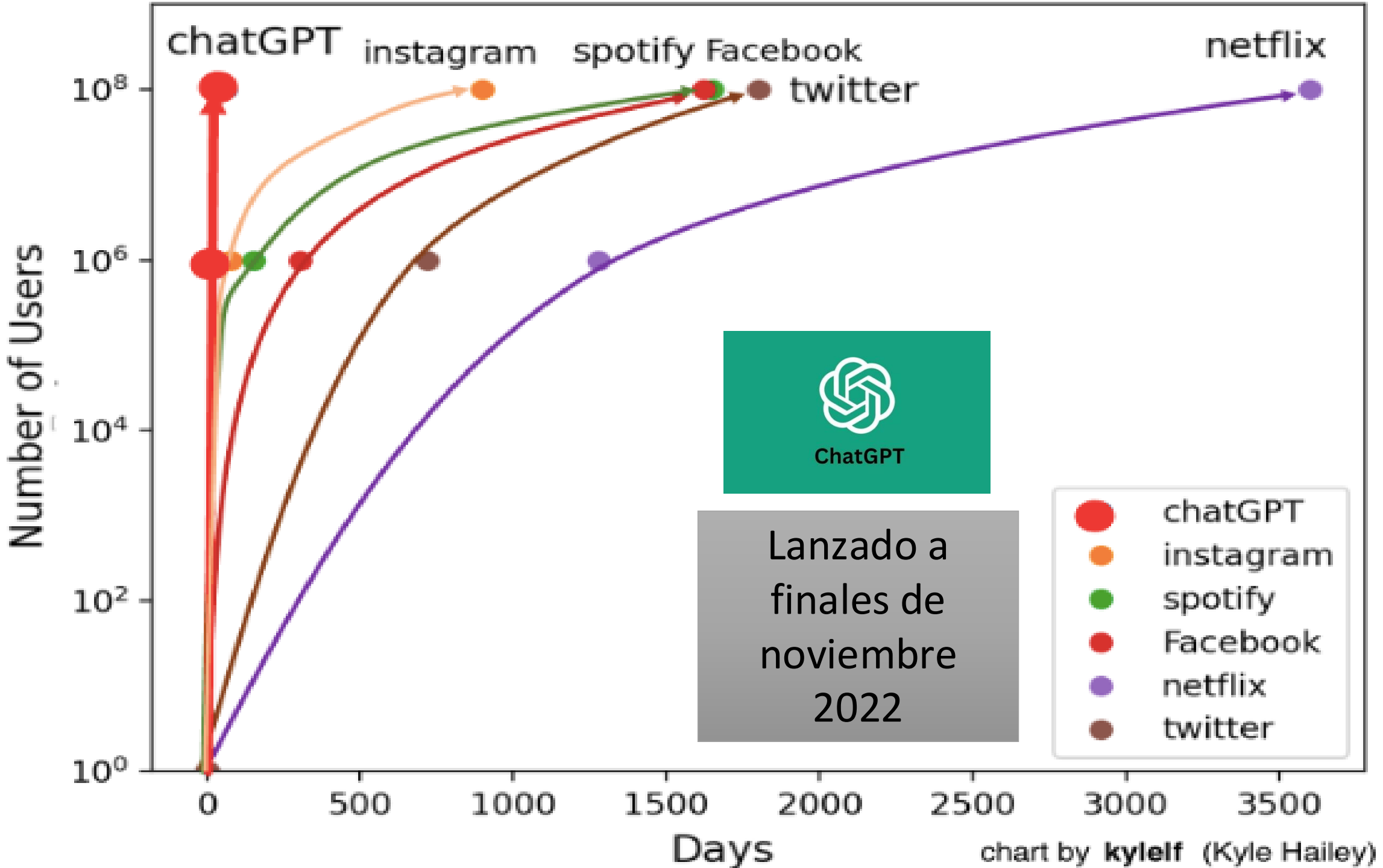
    - **Agrupación automática (Clustering):** Organizar fondos documentales agrupando contenidos que estén relacionados semánticamente.

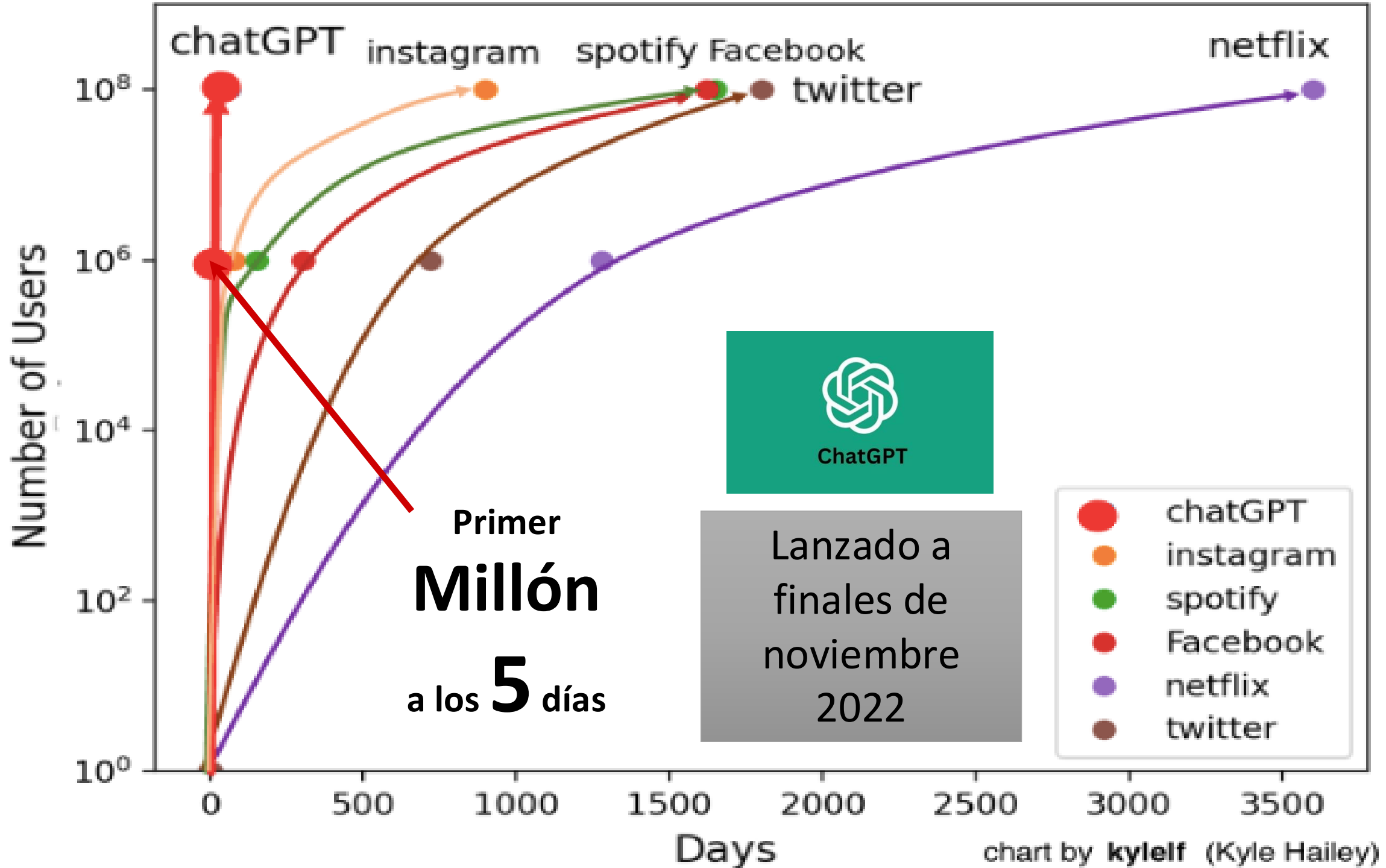


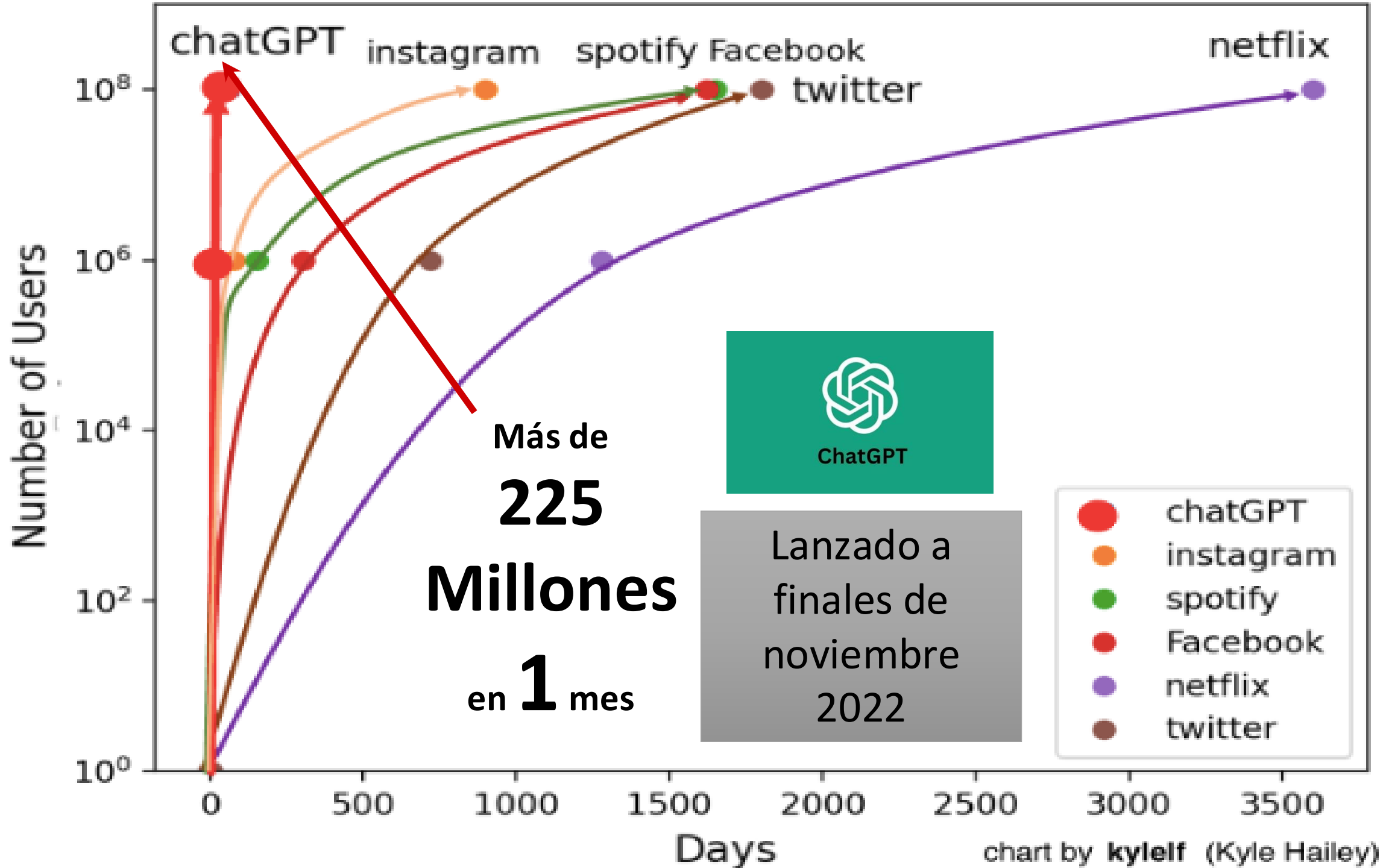
**Cursos Extraordinarios**
**Universidad de Zaragoza 2026**

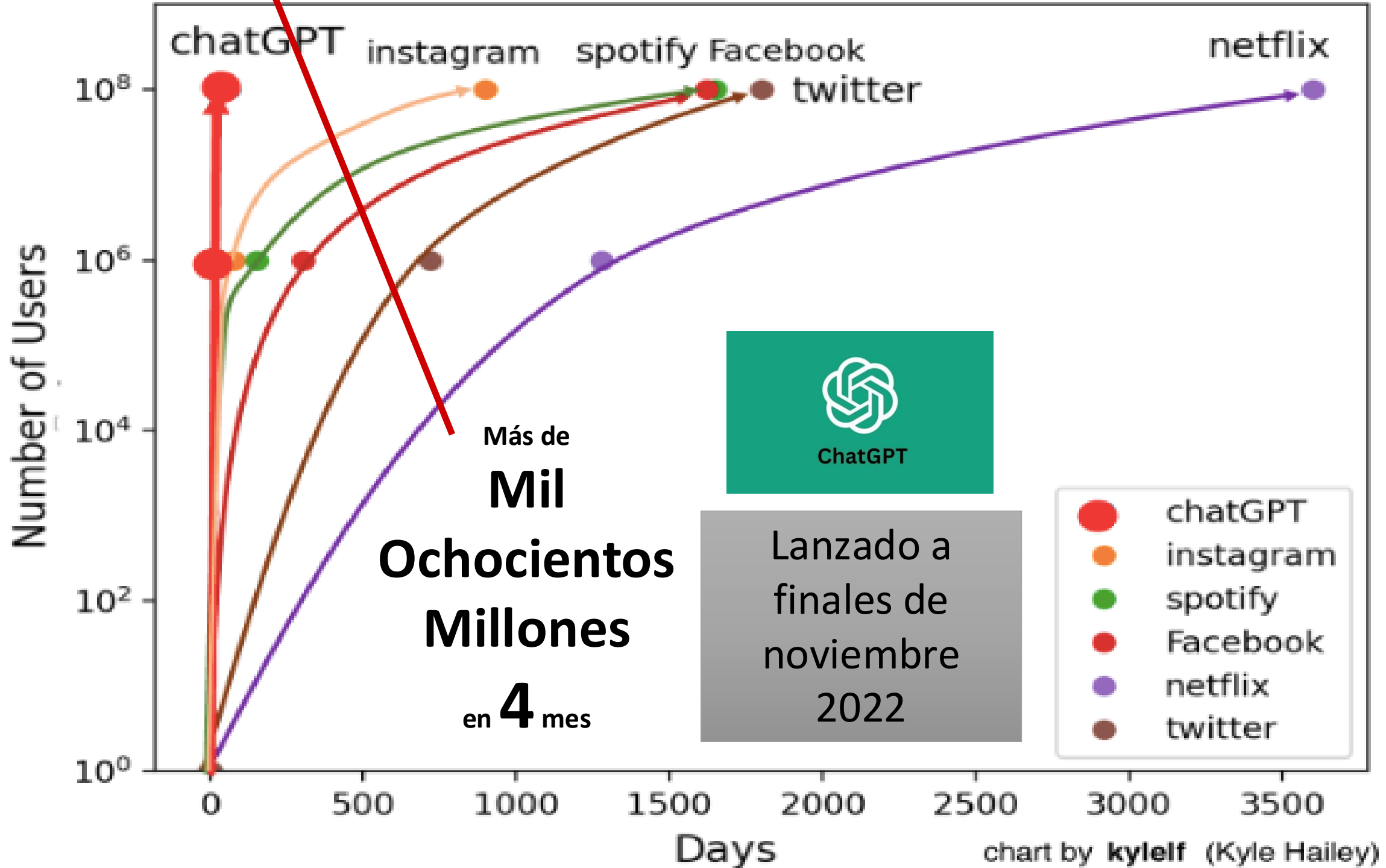
¿Qué tiene la IA para mi archivo?
De la teoría a la práctica

# Grandes Modelos de Lenguaje (LLM)

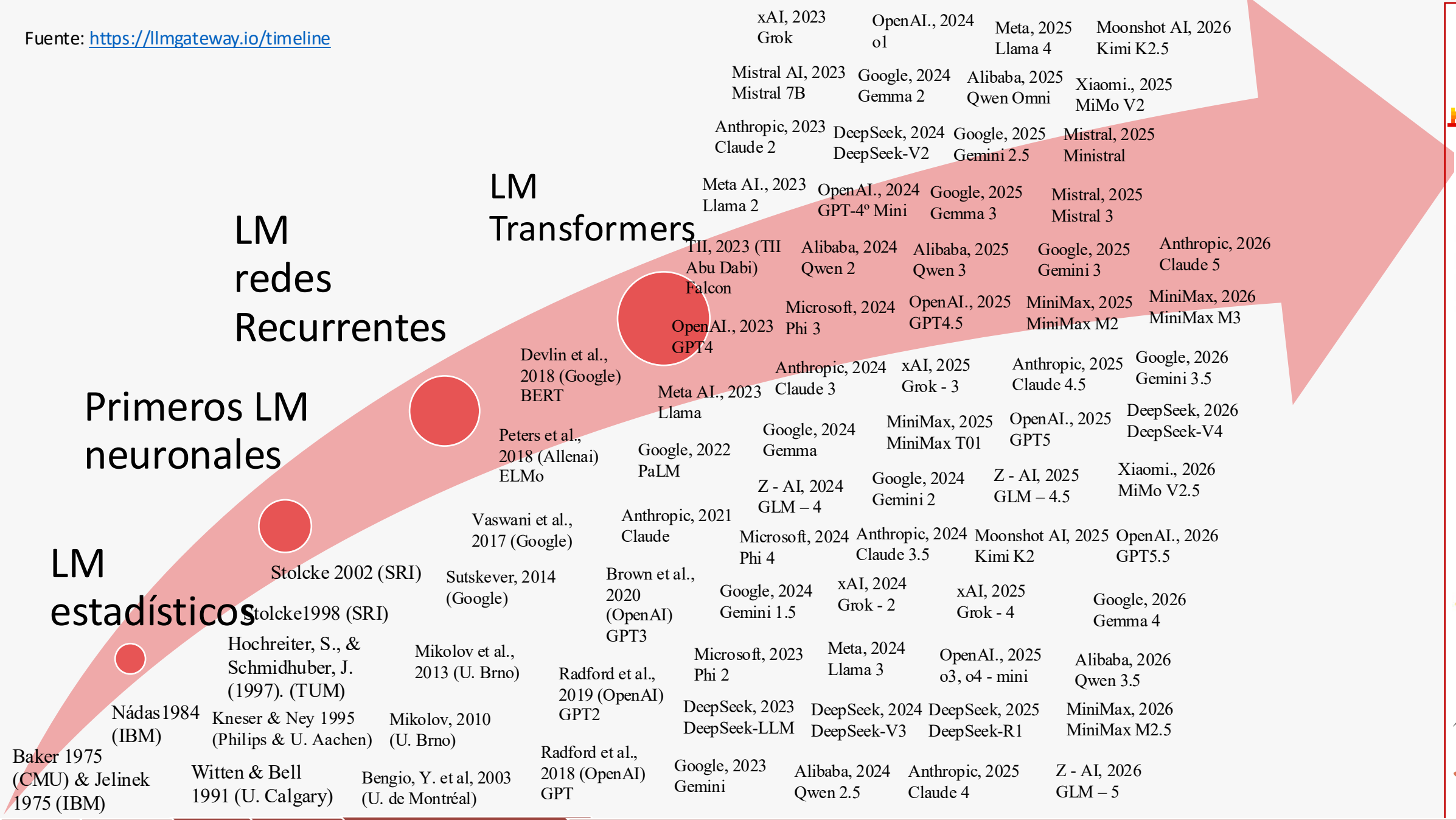

chatGPT
instagram
spotify
Facebook
twitter
netflix
Number of Users
$10^8$
$10^6$
$10^4$
$10^2$
$10^0$
0
500
1000
1500
2000
2500
3000
3500
Days
ChatGPT
Lanzado a finales de noviembre 2022
chatGPT
instagram
spotify
Facebook
netflix
twitter
chart by kylelf (Kyle Hailey)


chatGPT
instagram
spotify
Facebook
twitter
netflix
Number of Users
$10^8$
$10^6$
$10^4$
$10^2$
$10^0$
0
500
1000
1500
2000
2500
3000
3500
Days
ChatGPT
Primer
Millón
a los 5 días
Lanzado a finales de noviembre 2022
chatGPT
instagram
spotify
Facebook
netflix
twitter
chart by kylelf (Kyle Hailey)


chatGPT
instagram
spotify
Facebook
twitter
netflix
Number of Users
$10^8$
$10^6$
$10^4$
$10^2$
$10^0$
0
500
1000
1500
2000
2500
3000
3500
Days
Más de
225
Millones
en 1 mes
ChatGPT
Lanzado a finales de noviembre 2022
chatGPT
instagram
spotify
Facebook
netflix
twitter
chart by kylelf (Kyle Hailey)


chatGPT
instagram
spotify
Facebook
twitter
netflix
Number of Users
$10^8$
$10^6$
$10^4$
$10^2$
$10^0$
0
500
1000
1500
2000
2500
3000
3500
Days
chart by kylelf (Kyle Hailey)
Más de
Mil
Ochocientos
Millones
en 4 mes
ChatGPT
Lanzado a finales de noviembre 2022
chatGPT
instagram
spotify
Facebook
netflix
twitter

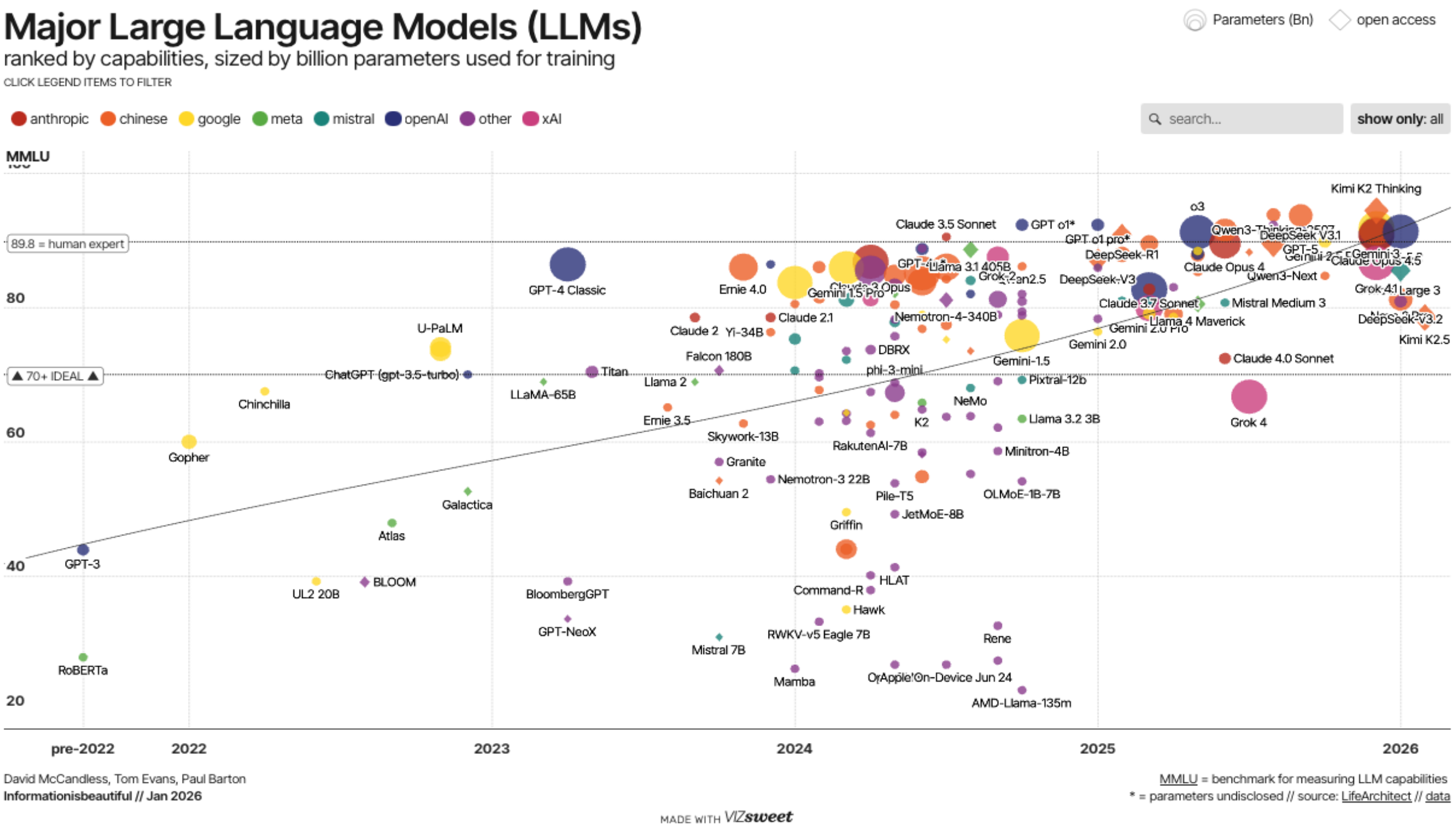
Fuente: https://llmgateway.io/timeline



# Major Large Language Models (LLMs)

ranked by capabilities, sized by billion parameters used for training

CLICK LEGEND ITEMS TO FILTER




David McCandless, Tom Evans, Paul Barton

**Informationisbeautiful // Jan 2026**

MMLU = benchmark for measuring LLM capabilities

* = parameters undisclosed // source: LifeArchitect // data

MADE WITH VIZsweet

Fuente: https://informationisbeautiful.net/visualizations/the-rise-of-generative-ai-large-language-models-llms-like-chatgpt?trk=public_post_comment-text

# Transformers

- **De lo Secuencial a la Atención:**
  - **Redes Recurrentes**:
    - Como leer un libro y solo poder apuntar en un folio
  - **Attention is all you Need**:
    - Como un examen con apuntes
- **¿Qué significa Large?:**
  - De **Milllones** a **Miles de Millones** (Billions)
  - **Habilidades Emergente**
    - GPT no fue diseñado para saber química, programar o jugar al ajedrez.
    - El haber visto tantos datos le enseñó.

**Attention Is All You Need**


**Ashish Vaswani*** Google Brain avaswani@google.com

**Noam Shazeer*** Google Brain noam@google.com

**Niki Parmar*** Google Research nikip@google.com

**Jakob Uszkoreit*** Google Research usz@google.com

**Llion Jones*** Google Research llion@google.com

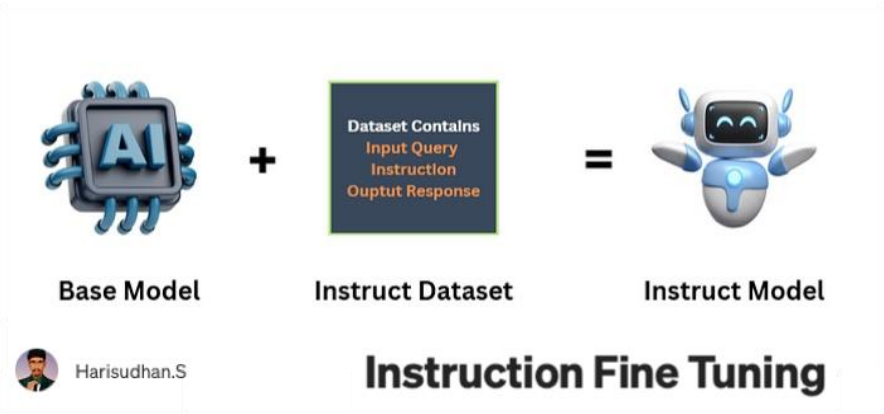
**Aidan N. Gomez*** † University of Toronto aidan@cs.toronto.edu

**Łukasz Kaiser*** Google Brain lukaszkaiser@google.com

**Illia Polosukhin*** ‡ illia.polosukhin@gmail.com

# Tipos de LLMs

- **Base Models vs. Instruct Models (Chat):**
  - **Base**:
    - Entrenados en cantidades masivas de datos sin etiquetar.
    - Completan el texto (predicen la siguiente palabra)
  - **Instruct/Chat (RLHF):**
    - Ha sido entrenado por humanos para ser útil, seguir instrucciones y no ser tóxico.
- **Open Source vs. Closed Source:**
  - Llama (Meta) o Mistral vs GPT-4 o Claude
  - ¿Por qué importa en la empresa?
    - Privacidad vs Potencia




https://medium.com/@speaktoharisudhan/instruction-fine-tuning-9ed1e510c0eb


# Modelos Abiertos

## Llama 3.1 / 3.3 / 4

- → **Llama 3.1/3.3:** Hito de 405B con contexto de 128K.
- → **Llama 3.2:** Modelos "Edge" (1B, 3B) y visión nativa.
- → **Llama 4:** Optimización masiva en razonamiento lógico y agentes.

## Mistral / NeMo / Pixtral

- → **Mistral Large 2:** 123B; eficiencia comparable a GPT-4o.
- → **Mistral NeMo:** 12B optimizado para hardware de consumo (NVIDIA).
- → **Pixtral:** Integración nativa de visión en 12B.

## Gemma 2 / 3 / 4

- → **Gemma 3:** Multimodalidad nativa y contexto extendido a 128K.
- → **Gemma 4:** (2026) Rendimiento superior en tamaños medianos (31B).
- → Arquitectura basada en el stack tecnológico de Gemini.

## Phi-3.5 / Phi-4

- → **Phi-4:** Avance crítico en razonamiento (14B) bajo licencia MIT.
- → **Phi-3.5 MoE:** 42B con 16B activos; alta densidad de conocimiento.
- → Optimización extrema para despliegue On-Premise.

## DeepSeek V3 / R1

- → **DeepSeek-R1 (2025):** Revolución en razonamiento abierto (RL).
- → Rendimiento en matemáticas y código igualando a OpenAI o1.
- → Catalizador del cambio hacia modelos de "Pensamiento".

## Qwen 2.5 / 3 /OLMo

- → **Qwen 2.5:** El nuevo estándar en benchmarks de codificación.
- → **OLMo:** Modelos 100% abiertos (datos y pesos) para investigación académica.

# LLMs Basados en Transformers

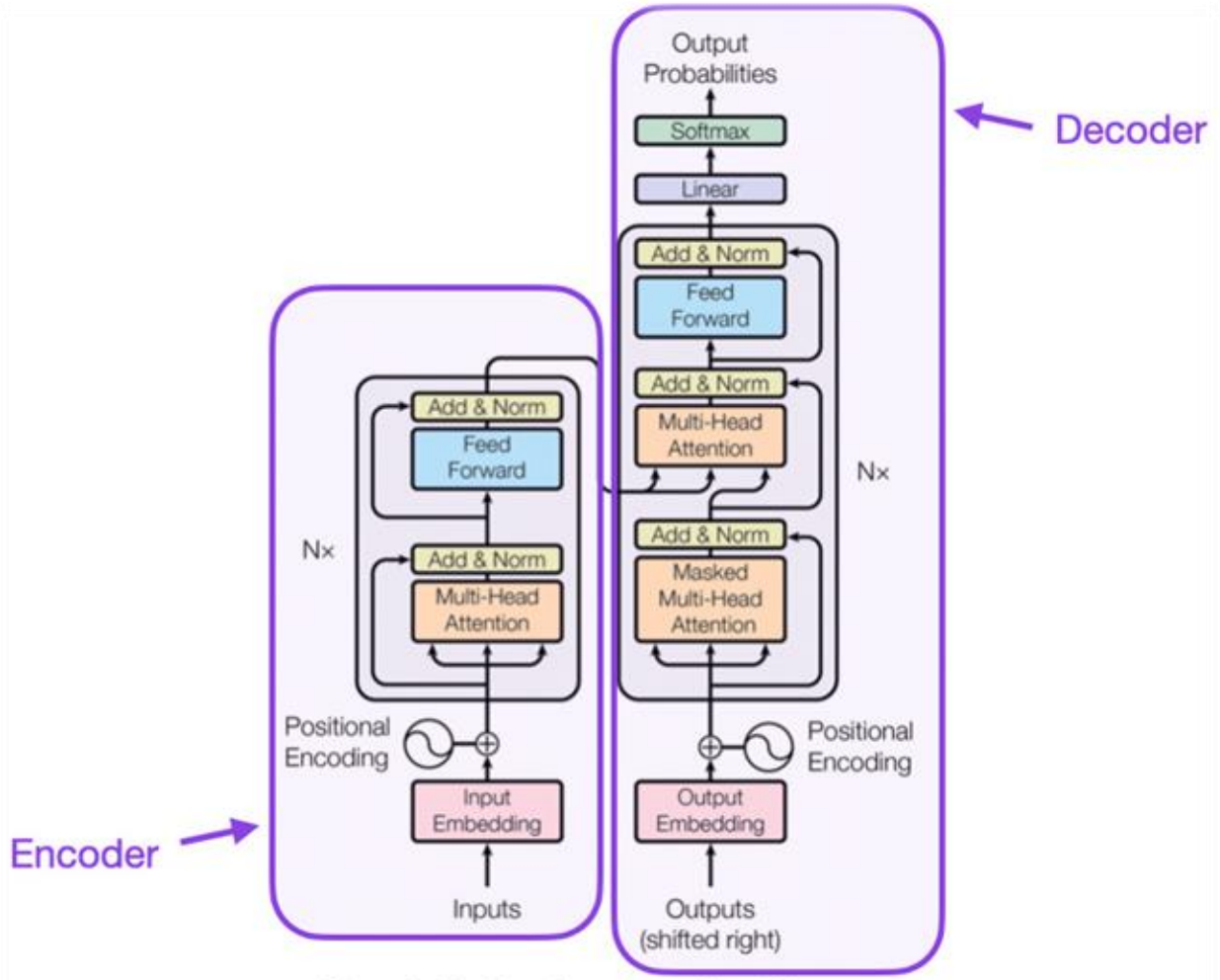



Figure 1: The Transformer - model architecture.

# LLMs Basados en Transformers

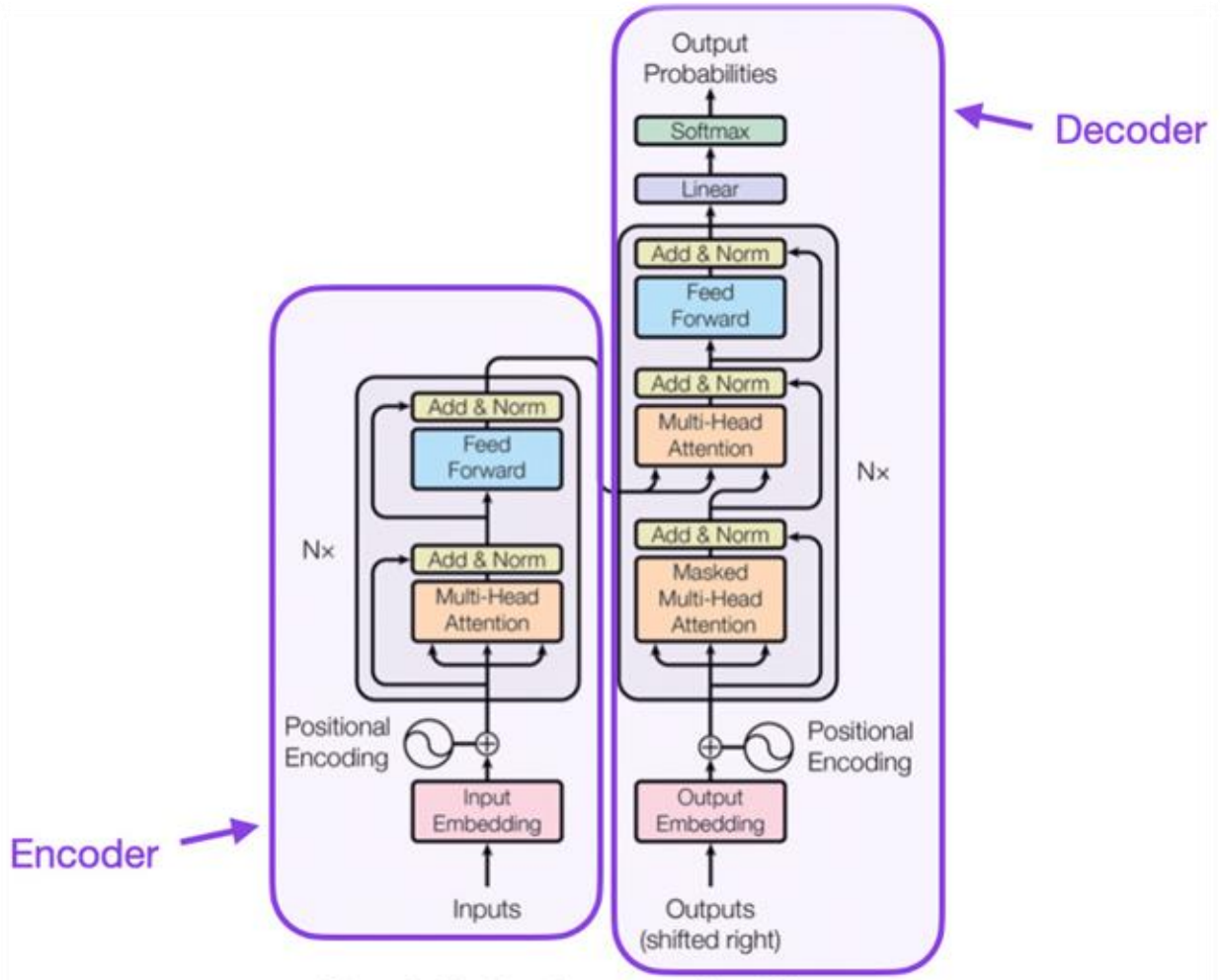



Figure 1: The Transformer - model architecture.

US011113602B2

(12) **United States Patent**
**Shazeer et al.**

(10) **Patent No.: US 11,113,602 B2**
(45) **Date of Patent: Sep. 7, 2021**

(54) **ATTENTION-BASED SEQUENCE TRANSDUCTION NEURAL NETWORKS**

(71) Applicant: **Google LLC**, Mountain View, CA (US)

(72) Inventors: **Noam M. Shazeer**, Palo Alto, CA (US); **Aidan Nicholas Gomez**, Toronto (CA); **Lukasz Mieczyslaw Kaiser**, Mountain View, CA (US); **Jakob D. Uszkoreit**, Portola Valley, CA (US); **Llion Owen Jones**, San Francisco, CA (US); **Niki J. Parmar**, Sunnyvale, CA (US); **Illia Polosukhin**, Mountain View, CA (US); **Ashish Teku Vaswani**, San Francisco, CA (US)

(73) Assignee: **Google LLC**, Mountain View, CA (US)

( * ) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **16/932,422**

(22) Filed: **Jul. 17, 2020**

(65) **Prior Publication Data**

US 2020/0372357 A1 Nov. 26, 2020

**Related U.S. Application Data**

(63) Continuation of application No. 16/559,392, filed on Sep. 3, 2019, now Pat. No. 10,719,764, which is a (Continued)

(51) **Int. Cl.**
***G06N 3/04*** (2006.01)
***G06N 3/08*** (2006.01)

(52) **U.S. Cl.**
CPC ..... ***G06N 3/08*** (2013.01); ***G06N 3/04*** (2013.01); ***G06N 3/0454*** (2013.01)

(58) **Field of Classification Search**
USPC ..... 706/15, 45
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS


2017/0124433 A1 5/2017 Chandrasekar et al.
2018/0341860 A1 11/2018 Shazeer
2020/0327359 A1* 10/2020 Bhandari ..... G06K 9/6215


FOREIGN PATENT DOCUMENTS


KR 1020150037986 4/2015


OTHER PUBLICATIONS


AU Examination Report in Australian Appln. No. 2018271931, dated Feb. 28, 2020, 3 pages.

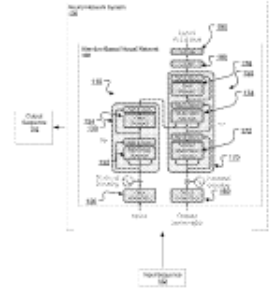

(Continued)

*Primary Examiner* — David R Vincent
(74) *Attorney, Agent, or Firm* — Fish & Richardson P.C.

(57) **ABSTRACT**


Methods, systems, and apparatus, including computer programs encoded on a computer storage medium, for generating an output sequence from an input sequence. In one aspect, one of the systems includes an encoder neural network configured to receive the input sequence and generate encoded representations of the network inputs, the encoder neural network comprising a sequence of one or more encoder subnetworks, each encoder subnetwork configured to receive a respective encoder subnetwork input for each of the input positions and to generate a respective subnetwork output for each of the input positions, and each encoder subnetwork comprising: an encoder self-attention sub-layer that is configured to receive the subnetwork input for each of the input positions and, for each particular input position in the input order: apply an attention mechanism over the encoder subnetwork inputs using one or more queries derived from the encoder subnetwork input at the particular input position.


**20 Claims, 3 Drawing Sheets**



# Los Planos:

El que Entiende / Lee

El que Genera / Escribe

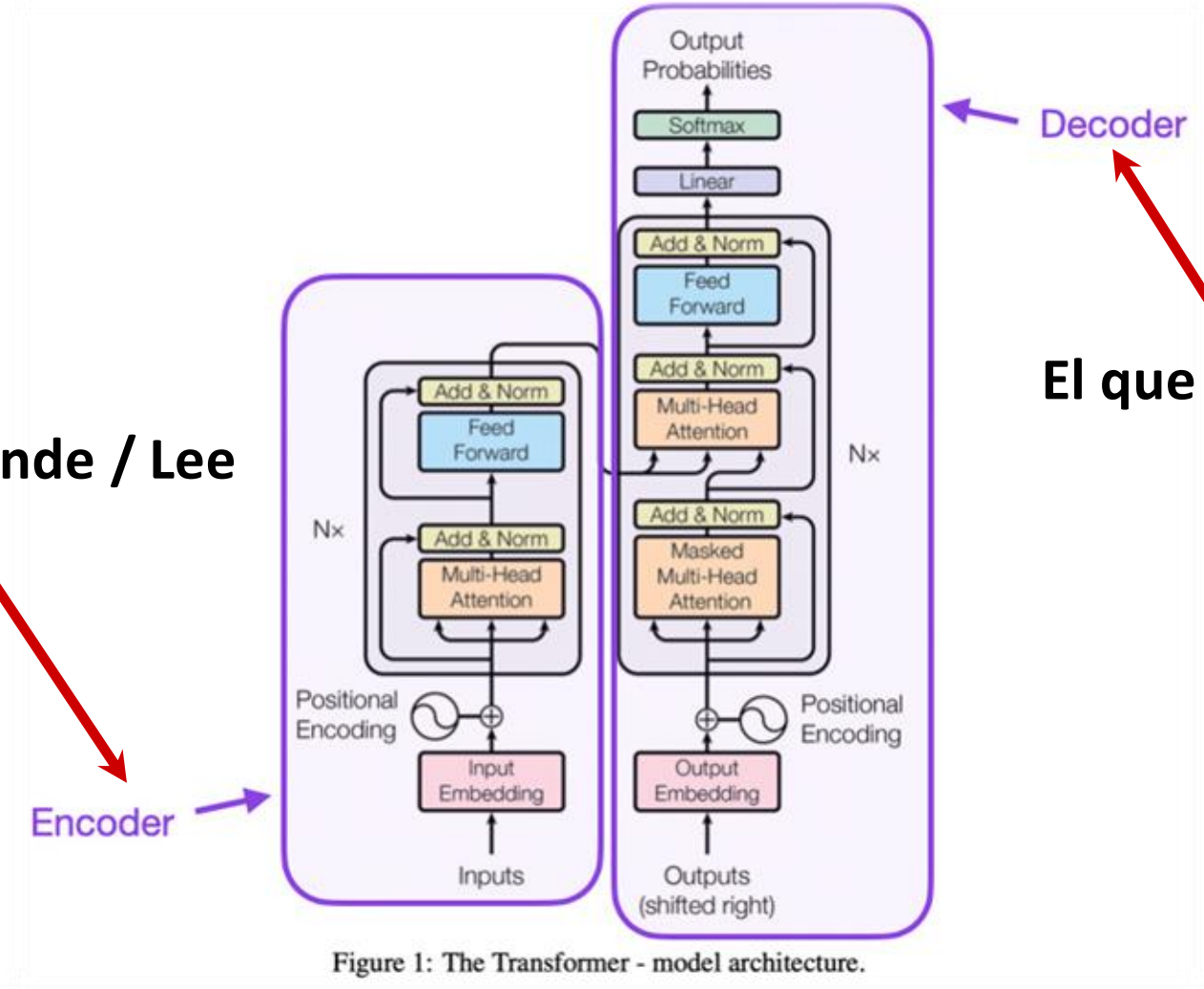
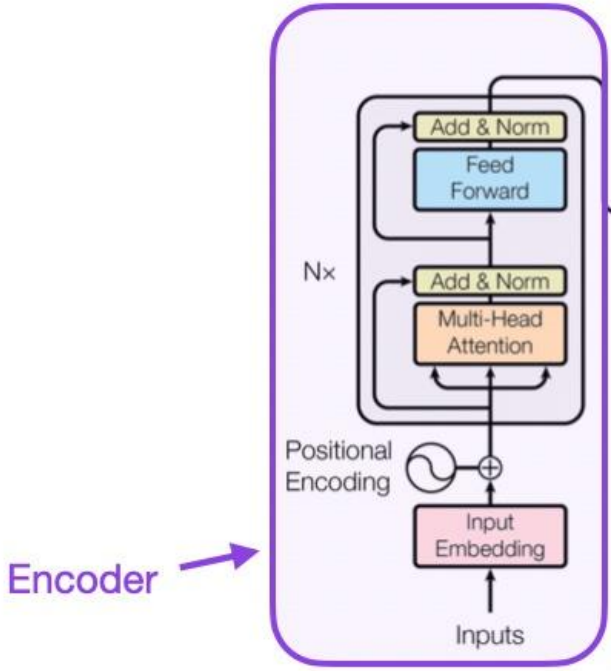


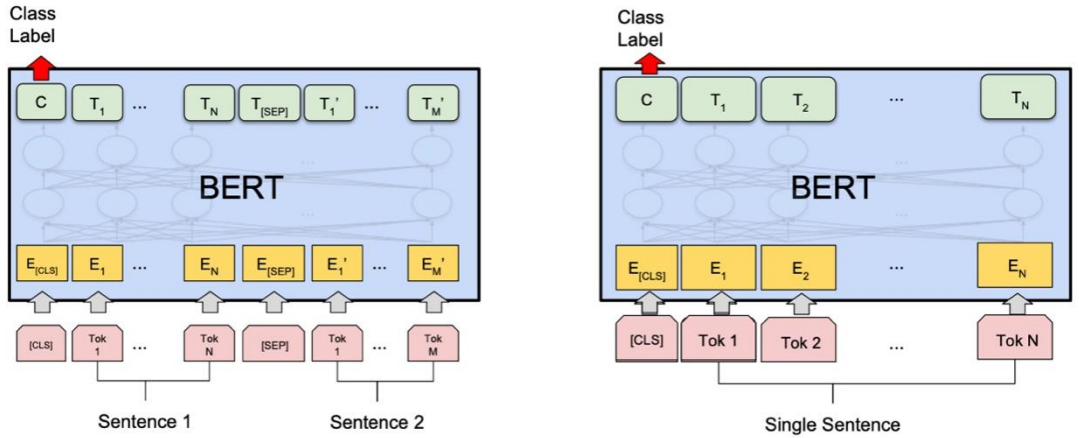

Figure 1: The Transformer - model architecture.

# Encoder Only: Análisis

- **BERT, RoBerta,…**




Encoder-style BERT model for predictive modeling tasks




(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

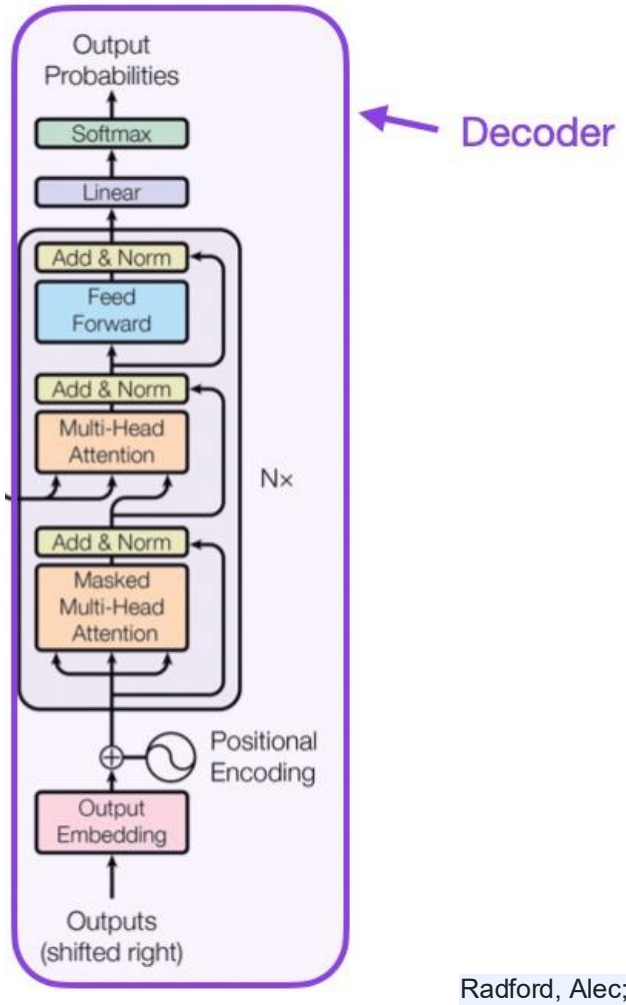
(b) Single Sentence Classification Tasks: SST-2, CoLA

- **Modelan** la entrada para tareas como la clasificación de texto
- Miran toda la frase a la vez: **Bidireccionales**:
  - Ven el contexto pasado y futuro de cada palabra
- **No generan** texto nuevo.
- Su trabajo es **entender** y **comprimir** el significado del texto en una lista de números (vector o embedding).
- Fundamental cuando veamos **RAG o Agentes** ya que las fuentes deberán ser analizadas y ”comprimidas”

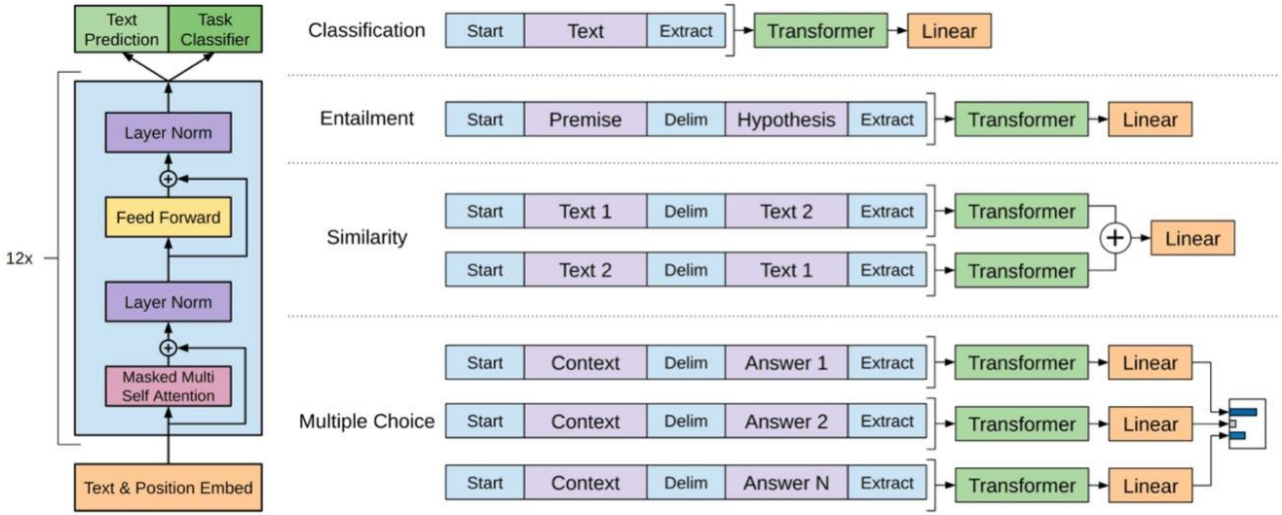
***BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*** **(2018)** by Devlin, Chang, Lee, and Toutanova https://arxiv.org/abs/1810.04805
**RoBERTa: A Robustly Optimized BERT Pretraining Approach** by Liu, Ott, et al. (2019) https://arxiv.org/pdf/1907.11692

# Decoder Only: Síntesis / Generación

GPT




Decoder-style GPT model (originally for predictive modeling)




- **Autorregresivo:** ”Predicen el futuro”.
- Toman lo anterior y **calculan la siguiente palabra** (token).
- No ‘ven' el futuro de la frase, **tienen que ir paso a paso**.
- **Dilema**:
  - Parece contraintuitivo. ¿Cómo puede ser tan inteligente algo que solo adivina la siguiente palabra?
- **Realidad**:
  - Si se escala el proceso lo suficiente (con billones de parámetros), para predecir bien la siguiente palabra necesitas entender lógica, contexto, ironía y hechos. La generación se convierte en razonamiento

Radford, Alec; Narasimhan, Karthik; Salimans, Tim; Sutskever, Ilya (2018). "Improving Language Understanding by Generative Pre-Training" https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

# Encoder - Decoder: Transductor y Multimodalidad

- **BART, T5, …**

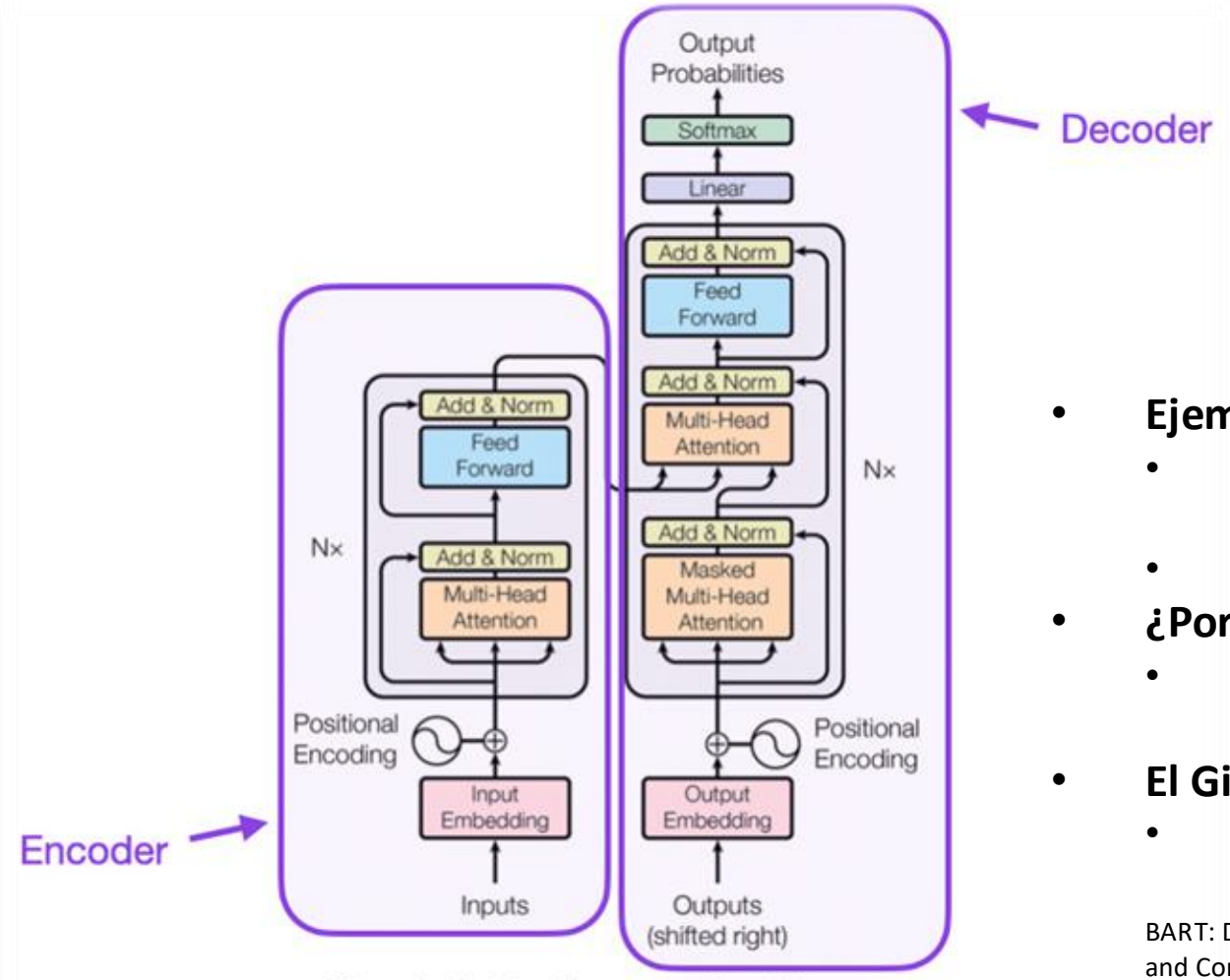
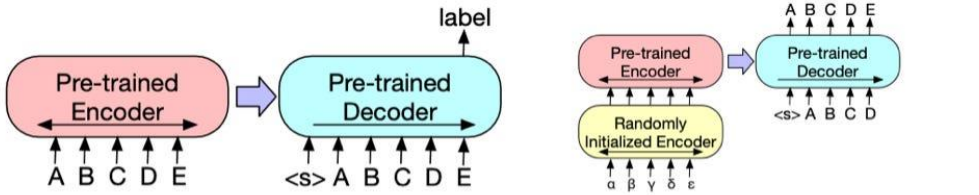



Figure 1: The Transformer - model architecture.

BART combines encoder and decoder parts




(a) To use BART for classification problems, the same input is fed into the encoder and decoder, and the representation from the final output is used.

(b) For machine translation, we learn a small additional encoder that replaces the word embeddings in BART. The new encoder can use a disjoint vocabulary.

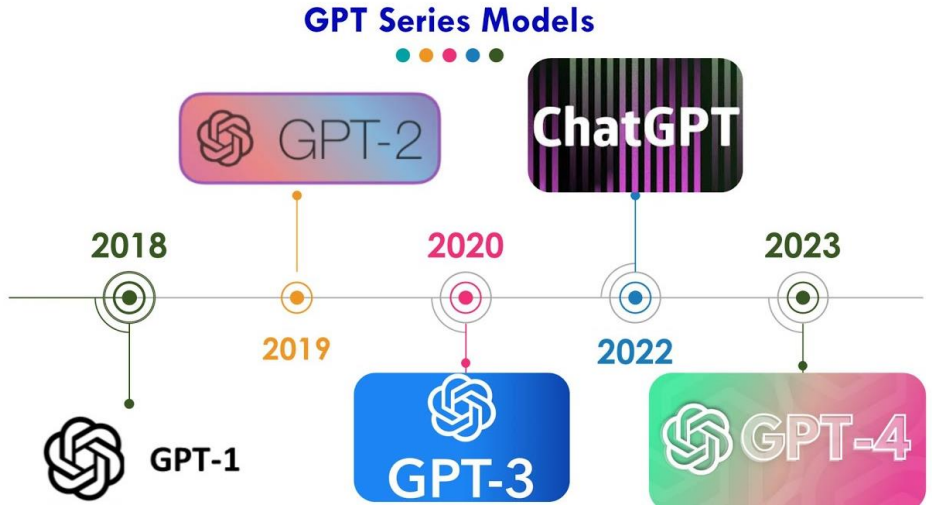
- **Ejemplos Clásicos:**
  - Traducción: Entra Inglés al Encoder -> sale Francés del Decoder
  - Resumen: Entra texto largo -> sale resumen sale.
- **¿Por qué se usan menos en chat?**
  - Porque los Decoder-Only son tan buenos que ya no hace falta el Encoder específico para texto.
- **El Giro Multimodal:**
  - Pero el Encoder-only no sirve cuando hay audio, imagen, …

BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension (2019) by Lewis, Liu, Goyal, Ghazvininejad, Mohamed, Levy, Stoyanov, and Zettlemoyer, https://arxiv.org/abs/1910.13461

# Resumen

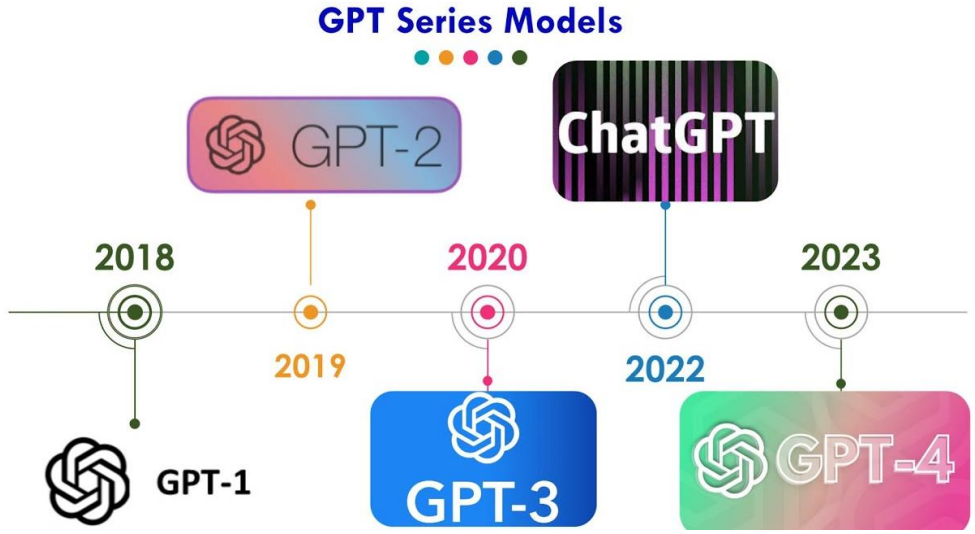
| Tipo | Estructura | Ejemplo Famoso | Rol / Analogía | Uso |
|---|---|---|---|---|
| **Encoder-Only** | Solo "Lee" | BERT, RoBERTa,… | Bibliotecario | Búsqueda semántica (RAG), Clasificación. |
| **Decoder-Only** | Solo "Escribe" | GPT-4, Llama, Gemini, Claude, Gemma, …. | Escritor / Improvisador | Chat, Redacción, Generación de ideas. |
| **Encoder-Decoder** | Lee y Transforma | T5, Whisper (Audio), Gemini (Multi) | Traductor / Ojos y Oídos | Traducción, Transcripción de reuniones, Análisis de imágenes. |

# Modelos Generativos




- **GPT-1 (2018): Pre-train + Fine-tune**
  - 117 Millones de parámetros
  - Pre "leyes de escala de los LLM"
  - Intento de resolver varias tareas de NLU: "**preentrenado** + **finetuning**"
    - **Pre-entrenado** primero con inmensas cantidades de texto
      - Para que **entienda** el lenguaje humano "en general", pero … con esto **no sabe hacer todavía nada**, si queremos realizar una tarea concreta: clasificar reseñas, contestar preguntas, resumir textos, etc. hay que "ajustarlo"
    - **Fine-tuning,** con datos específicos de la tarea y con etiquetas.
      - Con esto se aporta especialización.

# Modelos Generativos




- **GPT-2 (2019): Zero-shot incipiente**
  - 1500 Millones de parámetros

- **Multitarea** mediante **entrenamiento no supervisado** sin finetuning
- **Salida condicionada a la tarea**:
  - Salida dada la entrada, pero condicionada a la tarea
  - La tarea viene descrita en lenguaje natural con formato textual
- Prestaciones **todavía no superiores a modelos específicamente diseñados** para las tareas concretas: Q&A, comprensión, resumen, …
- Primeras muestras de capacidades **zero-shot** o **few-shot**:
  - Sin entrenarlo específicamente, podías pedirle (con suerte) que tradujera algo y a veces lo hacía.
  - El modelo empezaba a generalizar, pero era caótico e impredecible.

# Modelos Generativos

- **GPT-2 (2019): Zero-shot incipiente:**

System Prompt (human-written)

*In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.*

Model Completion (machine-written, 10 tries)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

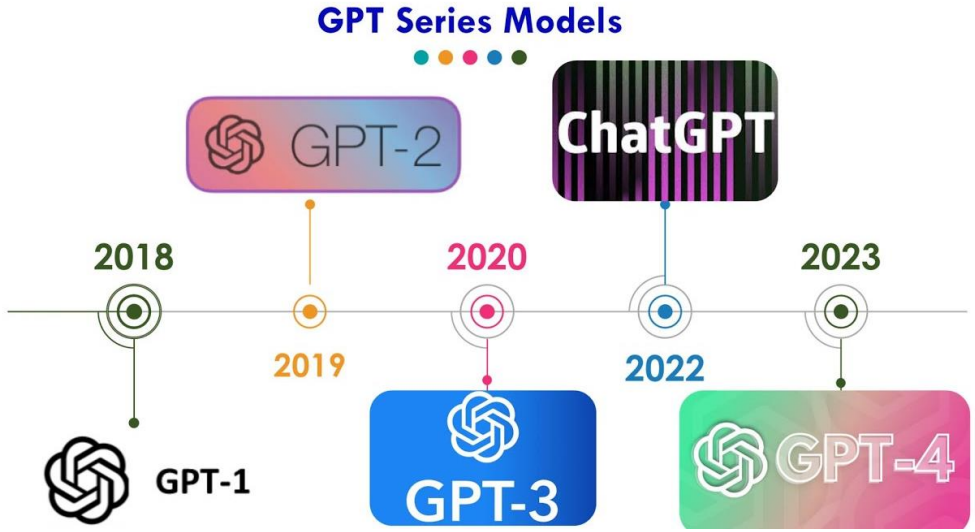
Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

# Modelos Generativos

- **GPT-3 (2020): La Revolución de la Escala (In-Context Learning)**




  - 175000 Millones de parámetros

  - Capacidad para desempeñar tareas (a partir de unos pocos ejemplos) para la que no fue entrenado:
    - "**task-agnostic**" & "**few-shot Learning**":
      - Es tan grande y ha visto tanto que no necesita re-entrenamiento.

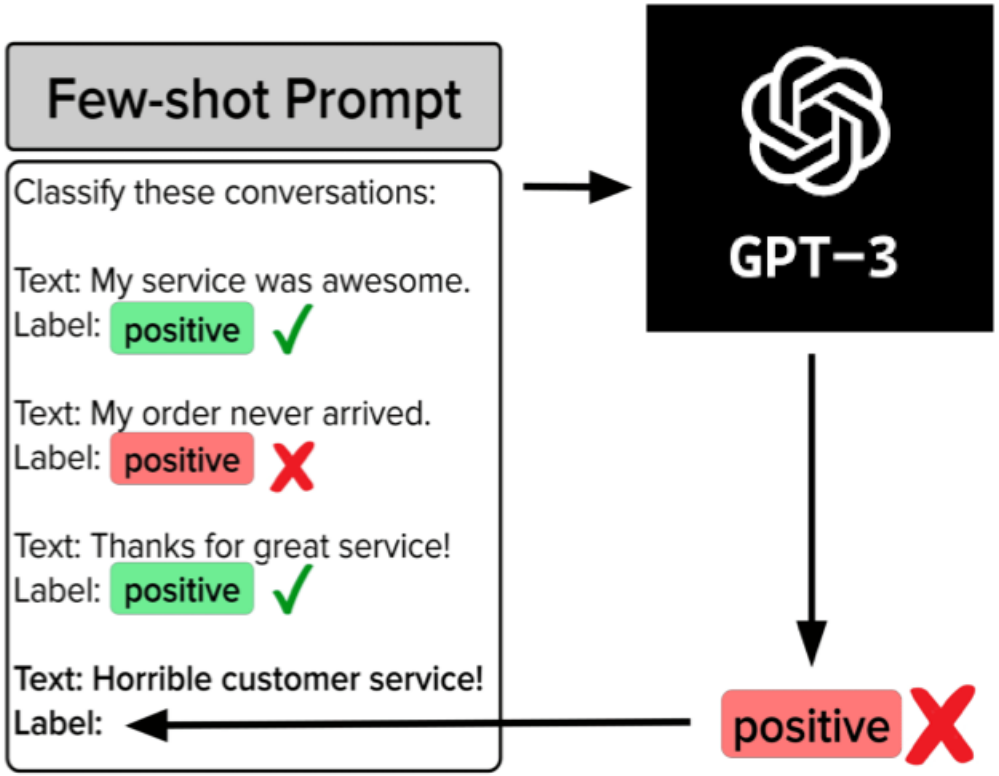
  - Competitivo frente a modelos específicamente entrenados para las tareas
  - Generación de textos indistinguibles de los textos humanos
  - Capacidad "**in-context Learning**" que permite tareas "zero-shot" o "few-shot":
    - El modelo infiere del contexto lo que debe hacer y responde en consecuencia

# Few-Shot Learning

- **Prompt:**
  - Inglés: Hello -> Español: Hola

    Inglés: Dog -> Español: Perro

    Inglés: Apple -> Español: ...
  - El modelo ve los ejemplos y dice: 'Ah, estamos jugando a traducir'.
    - Y completa: **Manzana**.



# Few-Shot Learning

```
// Here are the 2 description:code pairs used to give GPT-3
some context for how to provide a response

// sample 1
description: a red button that says stop
code: <button style={{color: 'white', backgroundColor:
'red'}}>Stop</button>

//sample 2
description: a blue box that contains 3 yellow circles with
red borders
code: <div style={{backgroundColor: 'blue', padding: 20}}><div
style={{backgroundColor: 'yellow', border: '5px solid red',
borderRadius: '50%', padding: 20, width: 100, height: 100}}>
</div><div style={{backgroundColor: 'yellow', borderWidth: 1,
border: '5px solid red', borderRadius: '50%', padding: 20,
width: 100, height: 100}}></div><div style={{backgroundColor:
'yellow', border: '5px solid red', borderRadius: '50%',
padding: 20, width: 100, height: 100}}></div></div>
```

carbon


Generación de layouts con JSX

# Few-Shot Learning

| | |
|---|---|
| Context → | Final Exam with Answer Key<br>Instructions: Please carefully read the following passages. For each passage, you must identify which noun the pronoun marked in *bold* refers to.<br>=====<br>Passage: Mr. Moncrieff visited Chester's luxurious New York apartment, thinking that it belonged to his son Edward. The result was that Mr. Moncrieff has decided to cancel Edward's allowance on the ground that he no longer requires *his* financial support.<br>Question: In the passage above, what does the pronoun "*his*" refer to?<br>Answer: |
| Target Completion → | mr. moncrieff |

# Madurez (2021 – 2022)

- **Refinamientos GPT 3 sobre datos específicos en dos líneas:**
  - **Línea Evolutiva 1: La Lógica y el Código:**
    - **Codex (2021)** Fine-tune masivo de GPT-3 con repositorios de GitHub.
      - Descubrimiento Sorprendente:
        - Al enseñarle a programar, no solo aprendió a escribir software, mejoró su capacidad de razonamiento y matemáticas.
        - ¿Por qué? Porque el código es lógica pura.
          No admite ambigüedades. Tiene estructura, bucles, condicionales (if, else).
          Al aprender la estructura del código, aprendió a estructurar su 'pensamiento'.

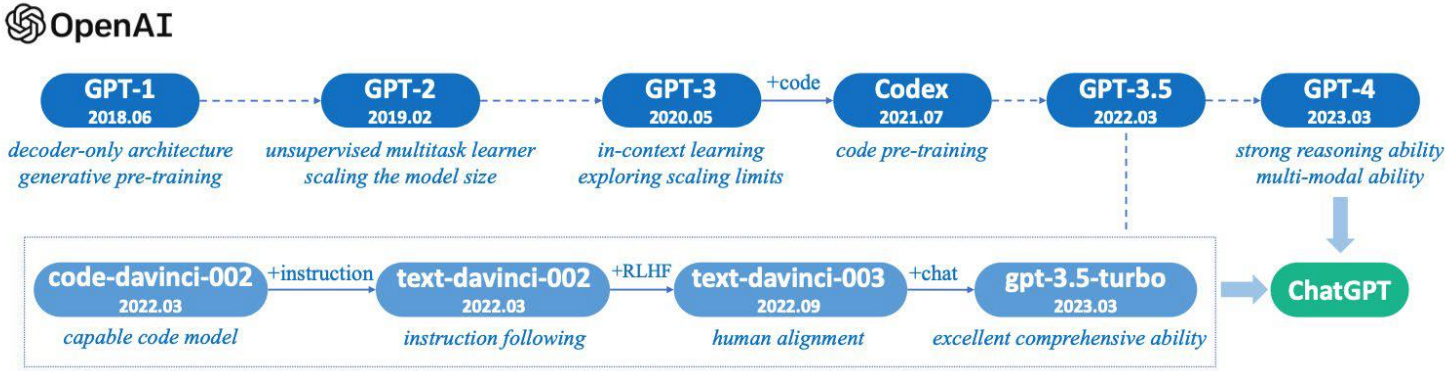
    - **GPT-3.5 code-davinci-002: Chain-of-Thoughts** (Cadena de Pensamiento):
      - Se fuerza que el modelo, al generar su respuesta, explique el "razonamiento" que lleva a la generación de dicha respuesta
      - Esto redujo las alucinaciones en tareas lógicas drásticamente."

# Madurez (2021 – 2022)

- **Línea Evolutiva 2: El Alineamiento Humano**
  - **Problema**: El modelo podía ser tóxico, racistas o no hacer caso a la instrucción.
    - Si le pedías 'Escribe una receta', a veces seguía con '...de la abuela que me gustaba mucho' en lugar de darte los ingredientes.
  - **Solución: InstructGPT (2022) y RLHF:**
    - **Paso 1: Supervised Fine-Tuning (SFT):**
      - Humanos escriben respuestas ideales para enseñar cómo se comporta un asistente útil.
    - **Paso 2: El Reward Model (Modelo de Recompensa):**
      - Se entrena otro modelo (Juez) para que aprenda qué es lo que gusta
        - Le damos al humano dos respuestas del modelo y le preguntamos: **'¿Cuál es mejor?'.**
        - Con miles de estas comparaciones, se crea un Reward Model capaz de predecir la preferencia humana."
    - **Paso 3: Optimización:**
      - Se usa ese **Reward Model** para **'puntuar'** al LLM:
        - Si es **tóxico** -> Puntos **negativos**. Si es **útil y seguro** -> Puntos **positivos**.
      - **Resultado**: Un modelo que no solo completa texto, sino que intenta **satisfacer la intención del usuario** de forma segura.

# Convergencia: ChatGPT (Finales 2022)

- **Unión de estas dos corrientes:**
  - **1: La capacidad de razonar y estructurar problemas:**
    - Heredada de Codex y el entrenamiento en código
  - **2: La capacidad de entender lo que queremos y ser seguro:**
    - Heredada de InstructGPT y RLHF
      - Conversaciones generadas por humanos en las que ambos actuaban como usuario y agente entre los datos de fine-tuning.
      - Mayor capacidad de comunicación con humanos
      - Amplio conocimiento intrínseco
      - Cierta capacidad de razonamiento para problemas matemáticos.

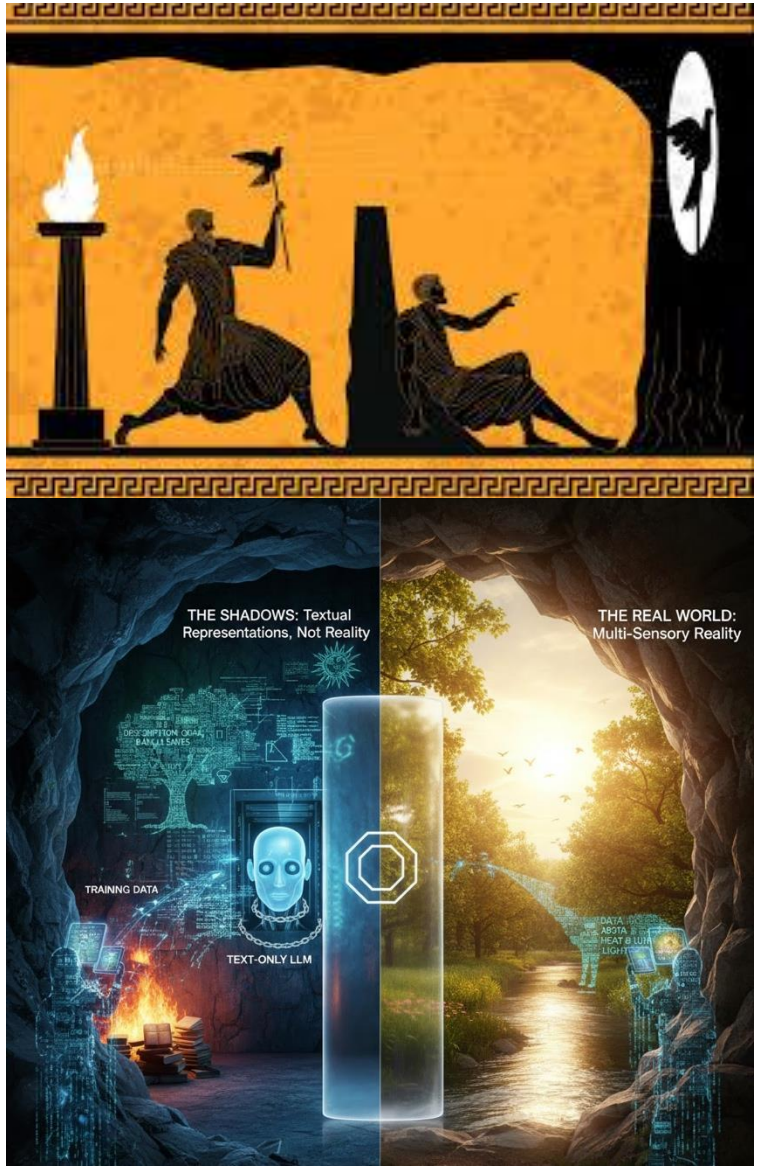


# Resumen

| Hito / Modelo | Año | Innovación Principal | Impacto |
|---|---|---|---|
| **GPT-3** | 2020 | Escala Masiva + Few-Shot | Capacidad generalista sin re-entrenamiento. |
| **Codex / Davinci-002** | 2021 | Entrenamiento en Código | **Razonamiento Lógico + Chain-of-Thought**. Resuelve problemas complejos paso a paso. |
| **InstructGPT** | 2022 | RLHF (Human Alignment) | **Seguridad y Obediencia**. Entiende instrucciones y evita toxicidad. |
| **ChatGPT** | 2022 | Chat Interface + Optimización | El paquete completo accesible para todos. |

# Despertando los Sentidos: La IA Multimodal

- **Modelos Solo Texto**
  - Como estar en una habitación cerrado conociendo el mundo solo a través de notas por debajo de la puerta
- **Multimodalidad**
  - Aportar información de los “sentidos”
    - Audio
    - Imagen
    - Video



# La Visión: CLIP y Contrastive Learning

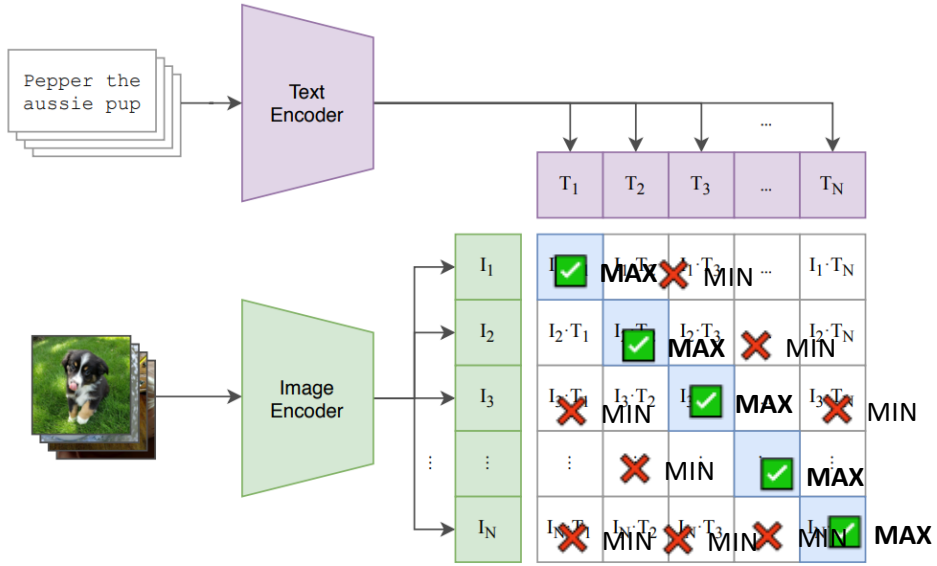
- **CLIP (Contrastive Language-Image Pre-training)**

**OpenAI (2021)**

- Aprendizaje Supervisado
  - Antes: Necesidad de datos etiquetados por humanos
    - Lento y caro
- Contrastive Learning
  - ¿Cómo funciona?
    - Como emparejar cartas:
      - Tenemos fotos y frases descriptivas (30k)
    - El modelo "acerca" matemáticamente cada foto con su descripción y la aleja de las demás.

# La Visión: CLIP y Contrastive Learning

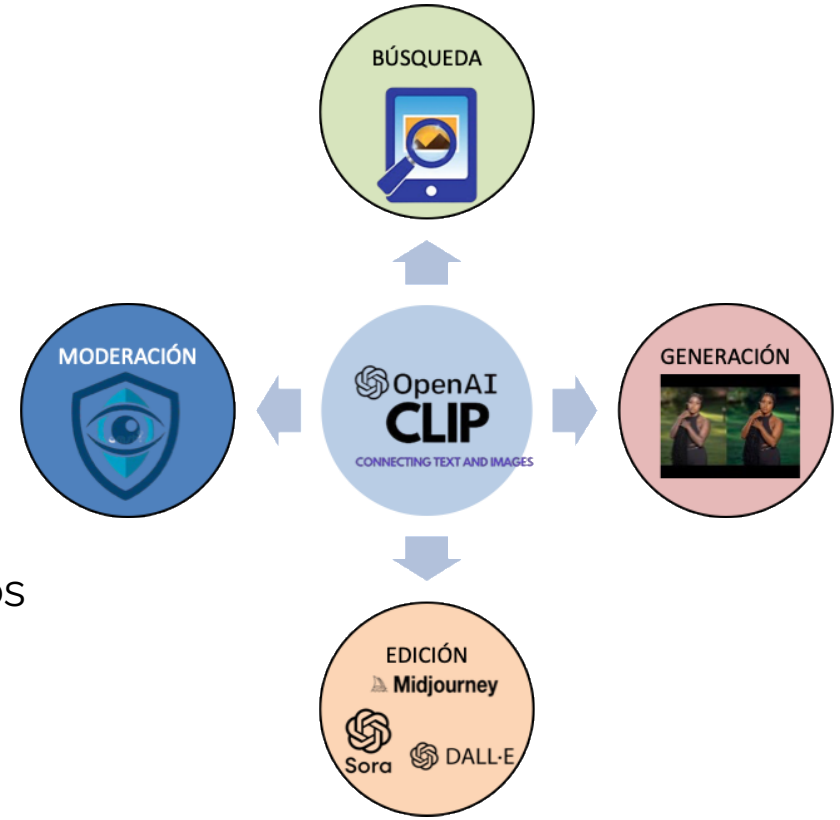
- **Internet es una mina de ORO**
  - **Contrastive Learning**
    - **2 Encoders**:
      Uno lee texto, el otro "lee" imágenes
    - El modelo "ve" pares **imagen-texto** sacados de Internet (**400 Millones**)




    - **Weak Supervision:**
      - Uso del "alt text" en HTML, pies de fotos, descripciones de productos, etc.
      - Si estaban juntos (foto y texto) es porque están relacionados:
    - Se comparan las representaciones de las fotos y los textos:
    - Se maximiza la cercanía entre la foto del perro y el texto "perro juando": P1– T1
    - Se minimiza la cercanía entre la foto del perro y la frase "plato de pasta": P1 - T17

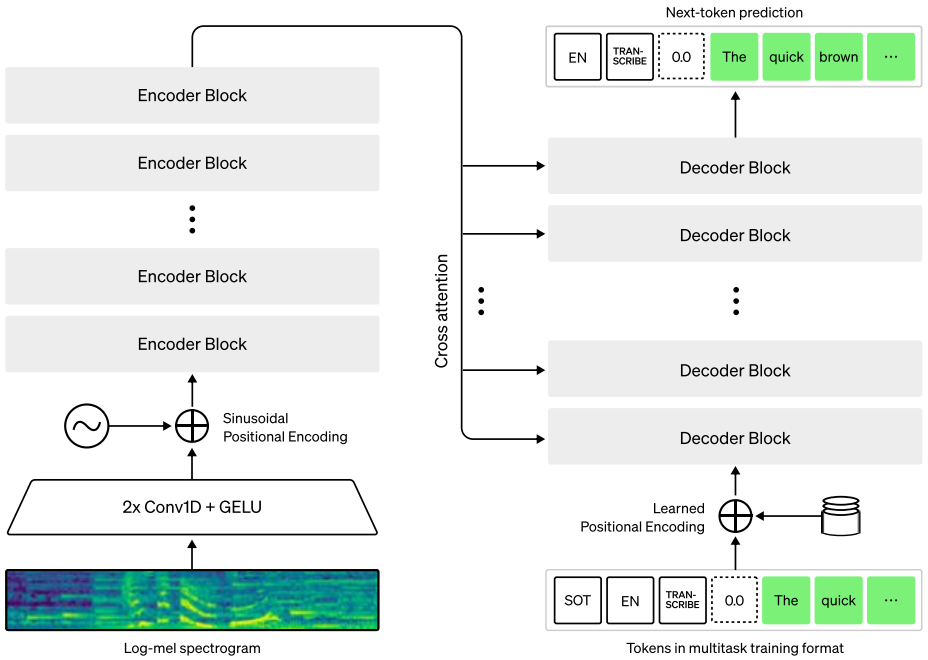
# La Visión: CLIP y Contrastive Learning

- **Legado de CLIP**
  - Catalizador de funcionalidades hoy comunes:
    - **Búsqueda** y Clasificación Visual Avanzada:
      - Marketing y Diseño: Con 50.000 fotos llamadas IMG_2023.jpg nadie encuentra nada.
      - Administración: Distinguir fotos de tickets y facturas, blletes de avión, …
    - **Generación** de Imagen a partir de Texto
      - Comunicación: Crear múltiples imágenes para posts, etc, …
      - Formación: Ilustraciones, diagramas o manuales para cursos
    - **Edición** y Manipulación Semántica de Imagen
      - Adaptar formatos y fondos, eliminar objetos distractores, ..
      - Extender una parte: horizonte, calle, … (outpainting)
    - **Detección** de Contenido y Moderación



# El Oído: Estructura Encoder-Decoder

- **Whisper (Open AI, 2022)**
  - Sustitución de reconocimiento automático del habla
  - Uso de estructura Encoder – Decoder clásico
    - ¡En lugar de codificar texto, codifica audio!
    - El decoder sigue prediciendo la siguiente palabra, pero condicionada por el audio
  - Entrenamiento masivo: Weak Supervision
    - 680.000 horas de audios
      - Modelo robusto (todo terreno)



| Modalidad | Tecnología Clave | Cómo funciona (Simplificado) | Uso |
|---|---|---|---|
| **Visión** | **CLIP** (Contrastive Learning) | Aprende a acercar imágenes y textos que significan lo mismo en un mapa matemático. | Búsqueda de imágenes, clasificación de productos, análisis de daños en seguros. |
| **Audio** | **Whisper** (Transformer) | Encoder-Decoder que transcribe voz a texto usando contexto. | Actas de reuniones, subtitulado automático, análisis de llamadas. |
| **Omni** | **GPT-4o / Gemini** | Un solo cerebro que procesa todo a la vez (Tokens de audio/video/texto). | Asistentes de voz en tiempo real, análisis de vídeo complejo. |

# IA Generativa Multimodal

- **Generación de Imagen:**
  - El modelo aprende a base de ”ver” un montón de imágenes borrosas, llenas de ruido
  - Se le pide que ”adivine” qué imagen hay debajo.
    - Al principio falla, pero al final aprenden los patrones “internos” de cómo está construida una imagen: bordes, sombras, texturas, …
  - En la práctica:
    - Prompt: "Un astronauta a caballo”
      - Análisis gracias a CLIP
      - Crea un “lienzo” lleno de ruido
      - Va “eliminando” ruido paso a paso hasta que aparece la imagen.
    - DALL-E 3, Midjourney, Stable Diffusion, Sora, Nano Banana Pro…

# LLM como asistente en la creación de prompts:

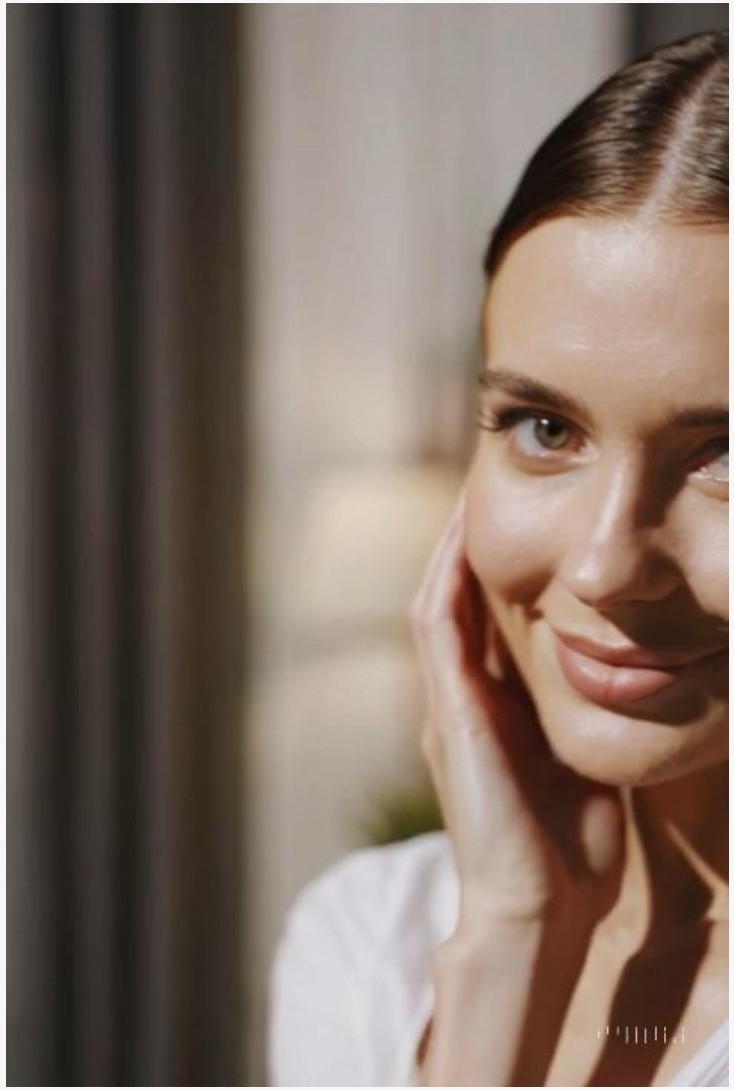
- **Un retrato fotográfico fotorrealista y cinematográfico de un astronauta solitario montando a caballo.**
  - **Sujeto:** El astronauta lleva un traje de la era Apolo muy desgastado y cubierto de polvo rojizo, con el visor del casco reflejando el paisaje. El caballo es un Quarter Horse musculoso de color castaño, con arreos de cuero antiguo y una manta de estilo western.
  - **Entorno:** Un vasto desierto similar a Monument Valley al atardecer. Formaciones de roca roja gigantes en el fondo.
  - **Atmósfera e Iluminación:** La "hora mágica" (golden hour). Luz solar rasante y cálida que crea largas sombras y un fuerte contraluz (rim lighting) alrededor de las figuras. Polvo en suspensión en el aire atrapando la luz.
  - **Detalles Técnicos:** Fotografía de película granulada (estilo Kodak Portra 400), tomada con una lente teleobjetivo de 85mm, profundidad de campo superficial (fondo desenfocado), colores ricos y texturizados. Composición épica de National Geographic.

# Generación de Audio

- El audio es diferente porque es secuencial
- **Voz (TTS Neural)**:
  - Modelos que crean representaciones espectrales de la voz (espectrogramas) y luego generan la onda.
  - Aprenden el contenido, y la prosodia.
- **Música (Suno, Udio)**:
  - Tratan la música como si fuera un lenguaje (estilo GPT).
  - Han escuchado toda la música del mundo y han aprendido qué nota, qué acorde y qué ritmo suele venir después de otro en cada estilo (Jazz, Rock, Clásica)….

# Generación de Video:

- El vídeo no es una imagen, son X imágenes por segundo que tienen que ser coherentes
- El Problema de la "Alucinación Temporal":
  - Si se generan imágenes por separado, en el segundo 1 la persona tiene camisa azul y en el segundo 1.1 la tiene roja. El vídeo parpadea y se deforma
- La Solución (Sora, Runway): Modelos de Difusión Espacio-Temporales
  - No se genera una imagen y luego la siguiente.
  - Se genera el bloque entero de vídeo de una sola vez, asegurándose de que las leyes de la física (gravedad, luz, permanencia de los objetos) se mantengan a lo largo del tiempo



| Tipo de Generación | ¿Qué hace? | Aplicación Real Inmediata |
|---|---|---|
| **Imagen** | Crea fotos, ilustraciones, bocetos desde texto. | Marketing rápido (RRSS), prototipado de producto, storyboards. |
| **Audio (Voz)** | Clona voces, lee texto con emoción humana. | Doblaje automático de formación, asistentes de voz empáticos. |
| **Audio (Música)** | Compone canciones o jingles desde cero. | Música de fondo libre de derechos para vídeos corporativos. |
| **Vídeo** | Crea clips de vídeo desde texto o anima fotos. | Anuncios conceptuales, visualización de productos en uso. |