

The background is a detailed investigation board. It features a central map with various colored lines (red, green, blue) connecting different points. Numerous handwritten notes on sticky papers are scattered across the board, some with question marks. There are also several small photographs of people, including a man in a suit and a woman. The notes contain various pieces of information, such as dates like 'EVA DAY 1', 'COLE DAY 2', and 'TAYLOR DAY 1', and questions like 'What was the motivation?' and 'Was Eva having an affair?'. The overall scene suggests a complex, multi-faceted investigation or a detailed project plan.

# Entendiendo ChatGPT:

## Funcionamiento, componentes clave y retos actuales de los Modelos de Lenguaje a Gran Escala

# Introducción

- Grandes Modelos de Lenguaje, Large Language Models (LLMs)
  - En meses han pasado de se ser uno de los modelos estrella de la última década a despertar **interés en la sociedad y medios**
  - ChatGPT fue lanzado el 30 de noviembre de 2022
    - En los meses siguientes hubo una reacción y número de usuarios inesperados

# Introducción

- **Grandes Modelos de Lenguaje, Large Language Models (LLMs)**
  - En meses han pasado de ser uno de los modelos estrella de la última década a despertar **interés en la sociedad y medios**
  - ChatGPT fue lanzado el 30 de noviembre de 2022
    - En los meses siguientes hubo una reacción y número de usuarios inesperados

The New York Times

## *The Brilliance and Weirdness of ChatGPT*

A new chatbot from OpenAI is inspiring awe, fear, stunts and attempts to circumvent its guardrails.

EL PAÍS

## “Funciona muy bien, pero no es magia”: así es ChatGPT, la nueva inteligencia artificial que supera límites

El modelo es la última evolución de los sistemas de generación de textos. Su coherencia sorprende a cientos de miles de usuarios con conversaciones creíbles que van desde derecho penal a lingüística

The Guardian

## ChatGPT reaches 100 million users two months after launch

Unprecedented take-up may make AI chatbot the fastest-growing consumer internet app ever, analysts say

# Introducción

- **Grandes Modelos de Lenguaje, Large Language Models (LLMs)**
  - En meses han pasado de ser uno de los modelos estrella de la última década a despertar **interés en la sociedad y medios**
  - ChatGPT fue lanzado el 30 de noviembre de 2022
    - En los meses siguientes hubo una reacción y número de usuarios inesperados

The New York Times

## *The Brilliance and Weirdness of ChatGPT*

A new chatbot from OpenAI is inspiring awe, fear, stunts and attempts to circumvent its guardrails.

EL PAÍS

## “Funciona muy bien, pero no es magia”: así es ChatGPT, la nueva inteligencia artificial que supera límites

El modelo es la última evolución de los sistemas de generación de textos. Su coherencia sorprende a cientos de miles de usuarios con conversaciones creíbles que van desde derecho penal a lingüística

The Guardian

## ChatGPT reaches 100 million users two months after launch

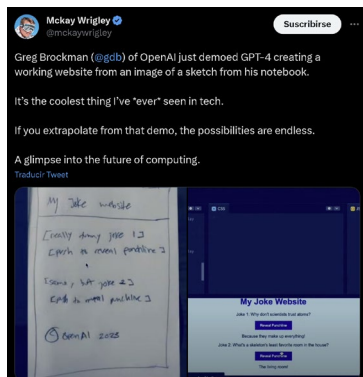
Unprecedented take-up may make AI chatbot the fastest-growing consumer internet app ever, analysts say

- Hoy trataremos de responder a algunas preguntas:
  - ¿Cómo empezó todo? ¿Qué son los LMs? ¿Cómo funcionan? ¿Quién los entrena? ¿Quién los adapta? ¿Son privados o públicos? ¿Son confiables? ¿Son seguros?... ¿Se pueden destilar?

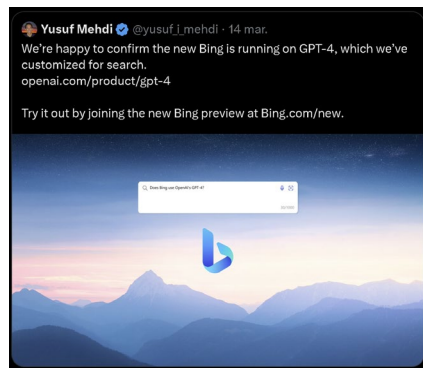


# Introducción

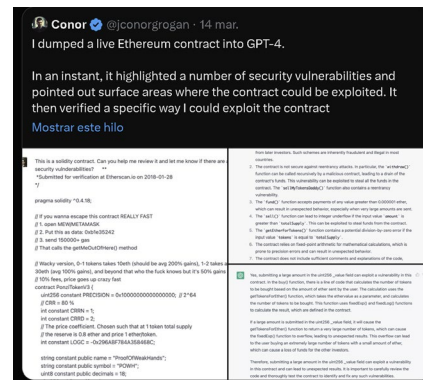
- GPT4 fue lanzado el 14 de marzo de 2023
  - en las primeras 24h se presentaron multitud de ejemplos de aplicaciones



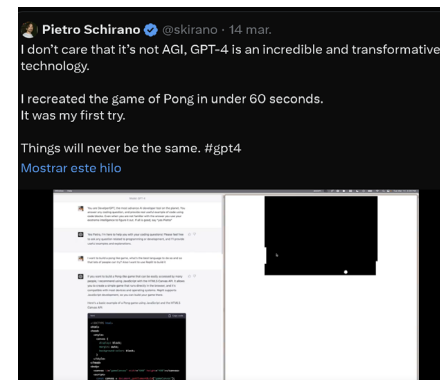
Asistencia desarrollo



Asistentes / búsquedas



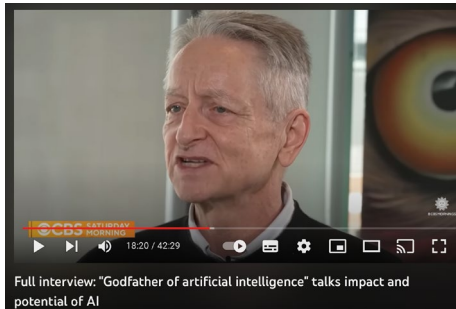
Revisión de contratos



Asistente programación

# Introducción

– Atención mediática a raíz de los LLMs: Investigadores y grandes empresas



02/05/2023

## Geoffrey Hinton / ex Google Brain

- "I don't think the technology we're building today naturally leads to artificial general intelligence,"
- "Until quite recently, I thought it was going to be like 20 to 50 years before we have general-purpose AI. And now I think **it may be 20 years or less**,"
- "We have to think hard about how you control that".



16/05/2023

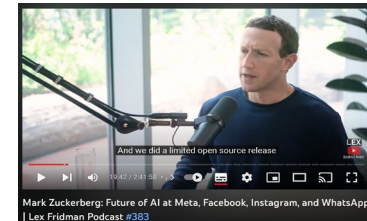
## Sam Altman / OpenAi

"I think if this technology goes wrong, it can go quite wrong."

"GPT-4 will, I think, entirely automate away some jobs, ... and it will create new ones that we believe will be much better."

"I think it's important to understand and think about GPT4 as a tool, not a creature,"

"We think the **regulatory intervention** by the governments will be critical to mitigate the risks of increasingly powerful models,"



11/06/2023

## Mark Zuckerberg/ Meta

Do you think the **LLAMA** or the language model underlying that version 2 will be open sourced? Do you have internal debate around that, the **pros and cons** and so on?



15/06/2023

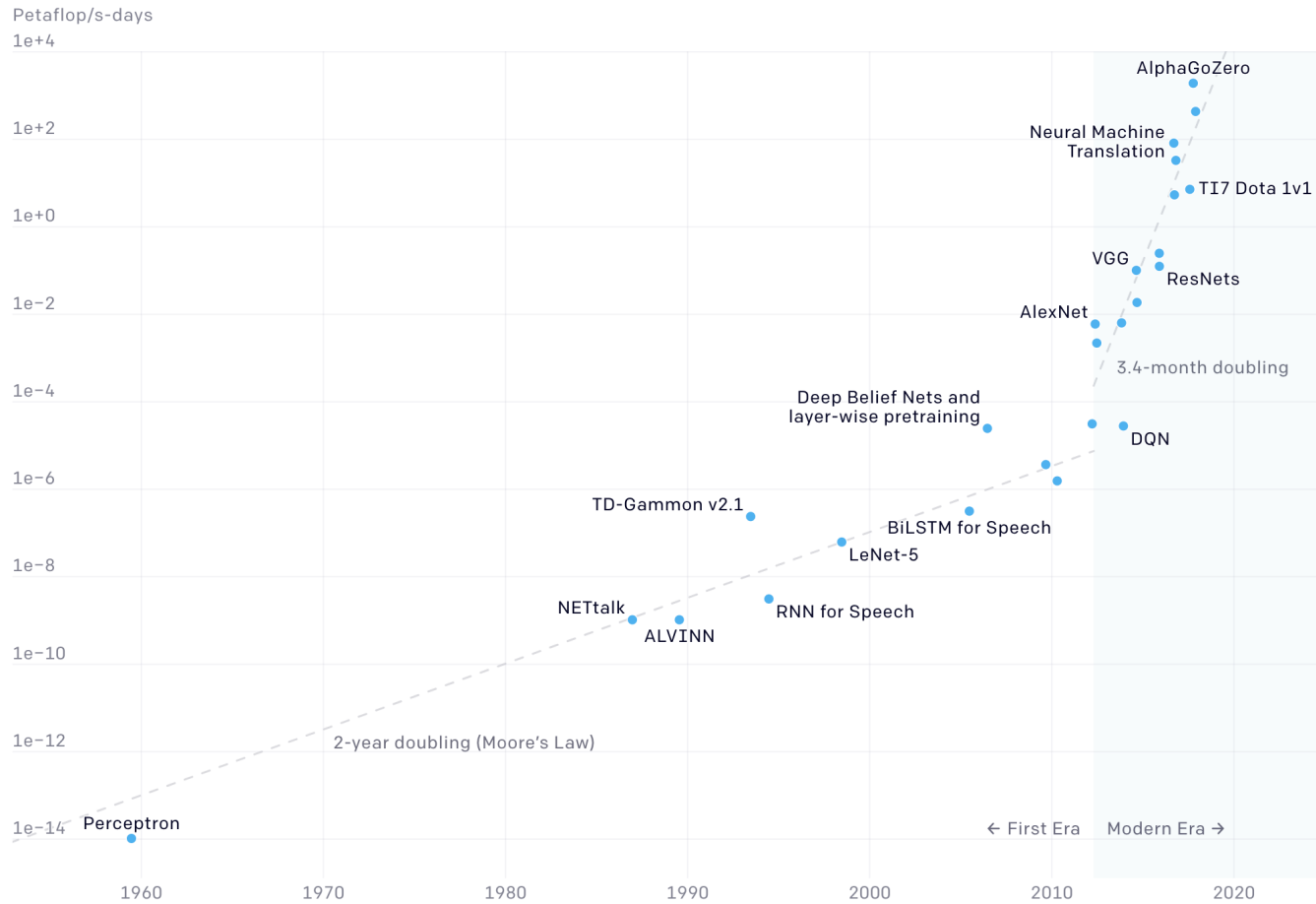
## Yann LeCun / Meta

it won't take over the world or permanently destroy **jobs**.

# Antecedentes

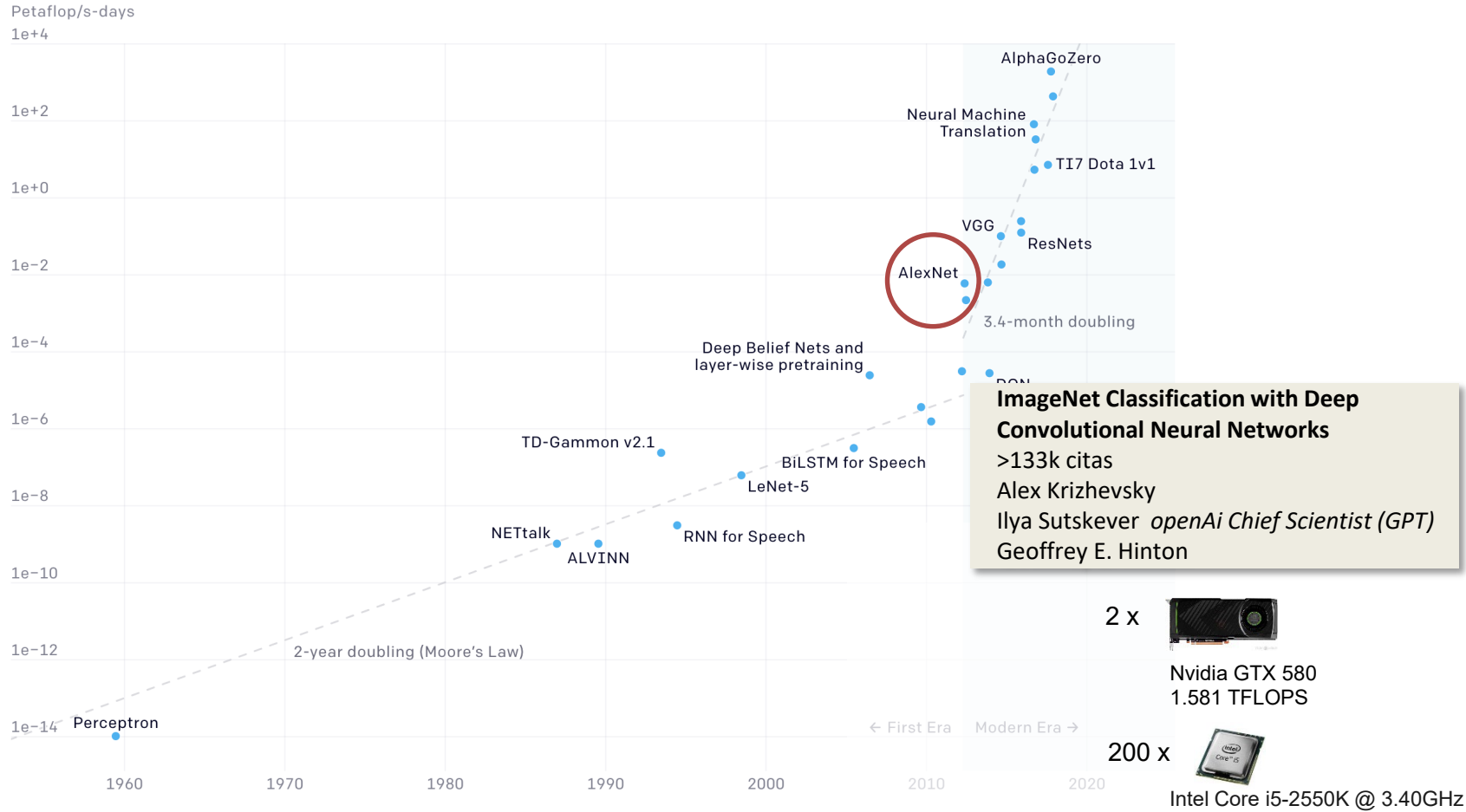
## Two Distinct Eras of Compute Usage in Training AI Systems

<https://openai.com/research/ai-and-compute>



# Antecedentes

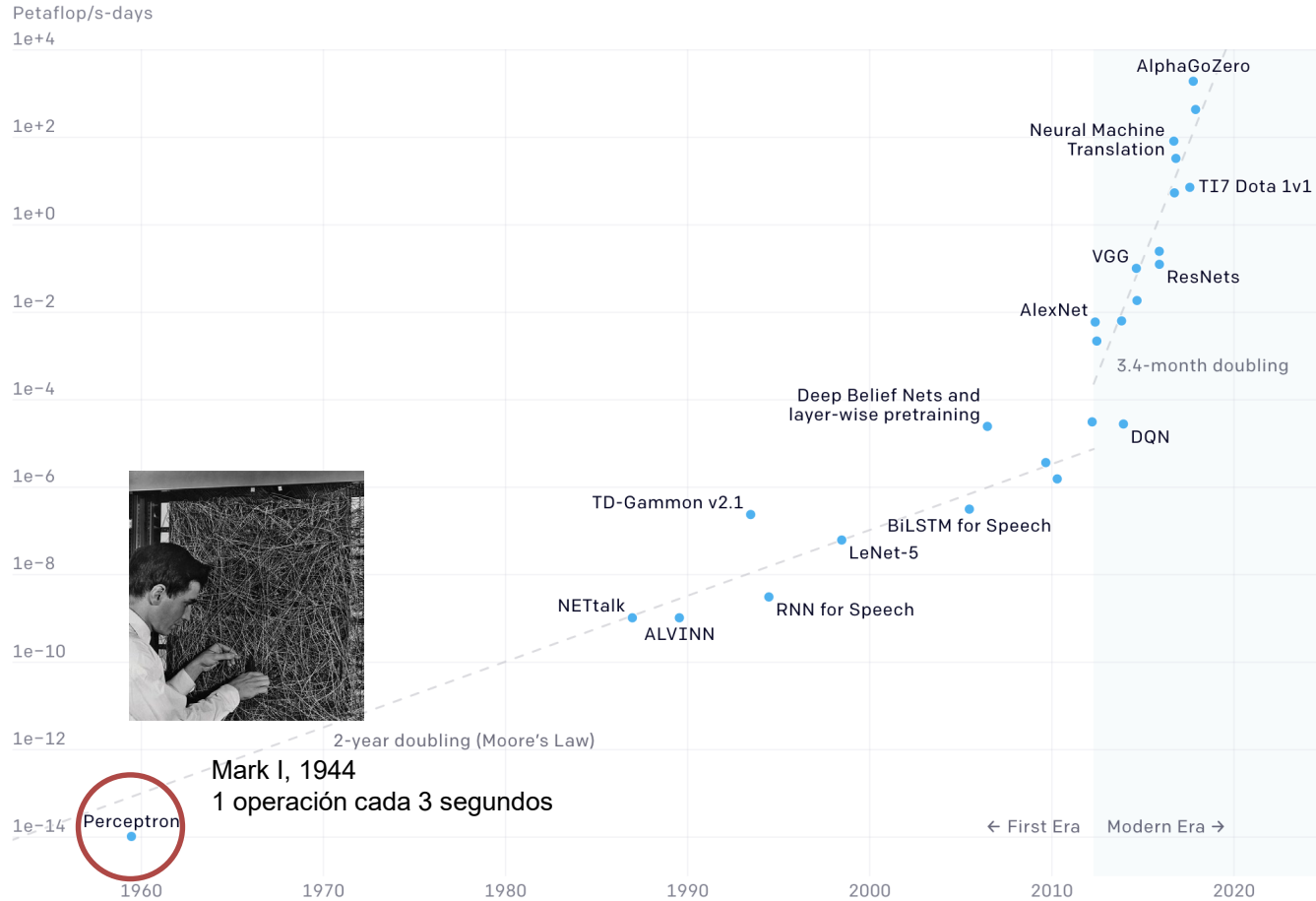
## Two Distinct Eras of Compute Usage in Training AI Systems





# Antecedentes

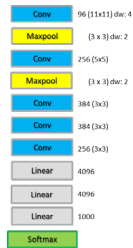
## Two Distinct Eras of Compute Usage in Training AI Systems



# Fundamentos

- Deep Learning: Ejemplo redes convolucionales

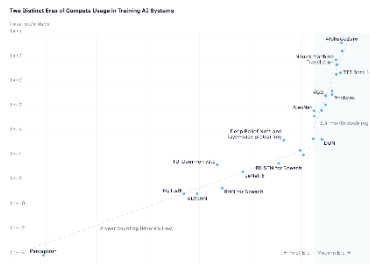
- A partir de 2012 se generó un gran interés por aumentar la profundidad de las redes (número de capas)



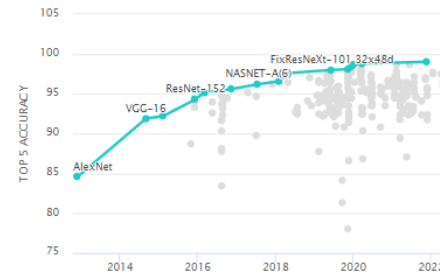
2012: **Alexnet** 7 capas  
84.6 % aciertos (Top 5)  
clasificación de imagen (Imagenet)

2014: **Inception** 25 capas  
93.3% aciertos

2015: **Resnet** >100 capas  
96.43% aciertos



recursos



rendimiento



# Fundamentos

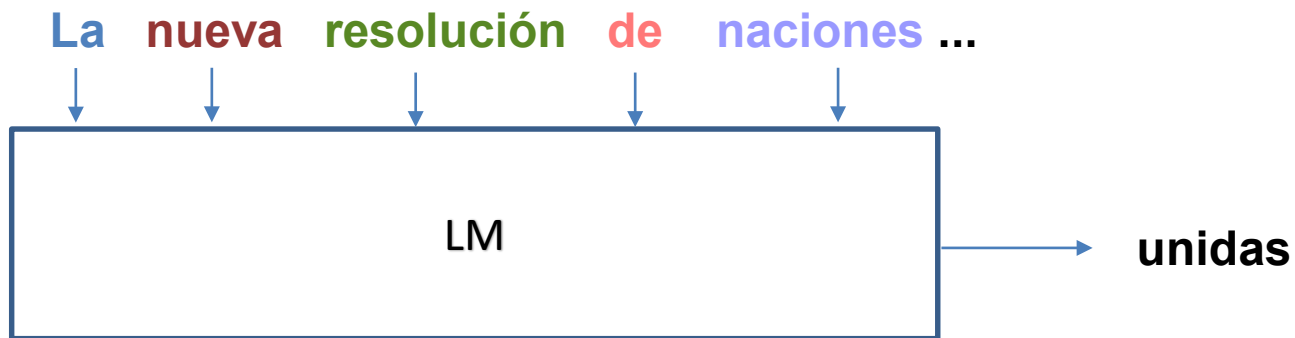
Modelos de lenguaje / Language models (LMs)

¿Qué son? ¿Para qué sirven? ¿Cómo funcionan?

# Fundamentos

- Modelo de lenguaje: objetivo

*predecir la siguiente palabra en una secuencia de palabras (un texto)*



# Fundamentos

- Modelo de lenguaje: partes de palabra (tokens)

**\_La** **\_nueva** **\_resol** **ución** **\_de** **\_n** **aciones...**



**Modelo LLAMA 7B**  
**32k tokens**

token		prob
----	----	----
<b>_un</b>		0.48
<b>_y</b>		0.04
<b>_en</b>		0.03
<b>_Un</b>		0.03
<b>_que</b>		0.02
<b>,</b>		0.02
<b>...</b>		

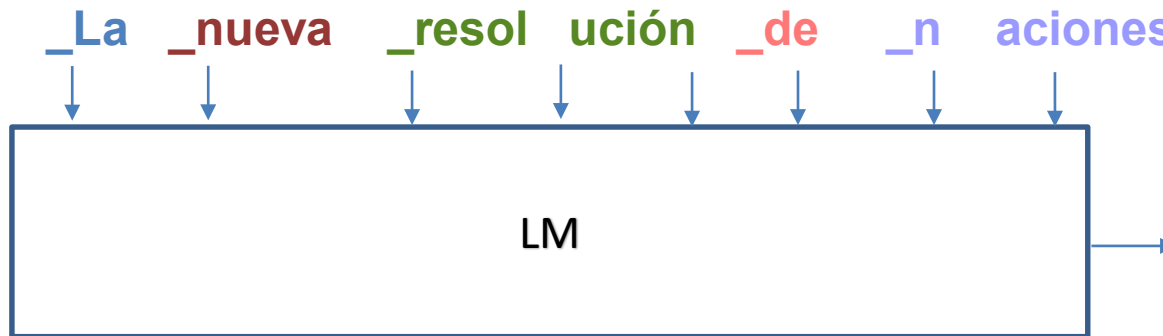
$p(\text{unidas} \mid \text{\_La, \_nueva, \_resol, ución, \_de, \_n, acciones})$

- Lo que obtendremos es una **estimación de la probabilidad**
- ¿Para qué queremos esas estimaciones?
- ¿No es suficiente con la más probable?



# Fundamentos

- Modelo de lenguaje: generación



## Modelo LLAMA 7B (mejores continuaciones)

```
-----
_un (0.48) ->
  idas (0.99)
  ific (0.00)
  ip (0.00)
  as (0.00)
  id (0.00)
-----
_y (0.04) ->
  _pue (0.21)
  _g (0.07)
  _raz (0.04)
  _nacional (0.04)
  _de (0.04)
-----
_en (0.03) ->
  _la (0.13)
  _el (0.13)
  _conflic (0.05)
  _las (0.04)
  _Europa (0.03)
```

- **Generar secuencias** (la salida de los LMs generativos)
- **Controlar otros procesos:** Reconocedor de automático del habla, traductores, etc.

# Fundamentos

- Representación de palabras

- Cada palabra del vocabulario se representa mediante un **vector** (32k por ejemplo en LLAMA)
- Pensemos en una versión simplificada donde cada palabra fuera descrita mediante una ficha de miles de preguntas (sparse vector)

palabra	rey	reina	gato	árbol	ráiz	gustará	odiaba
-----							
propiedades							
persona	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>					
monarquía	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>					
vegetal	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		
animal			<input checked="" type="checkbox"/>				
-----							
masculino	<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>			
femenino		<input checked="" type="checkbox"/>			<input checked="" type="checkbox"/>		
-----							
positivo						<input checked="" type="checkbox"/>	
negativo							<input checked="" type="checkbox"/>
-----							
verbo					<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>
presente							<input checked="" type="checkbox"/>
pasado							<input checked="" type="checkbox"/>
futuro					<input checked="" type="checkbox"/>		

Un cambio de significado puede ir asociado a cambiar uno de esos aspectos, **recuerdan más a un ideograma chino** que a una representación mediante un alfabeto fonético

木 → 本  
Árbol      Raíz

- palabras semánticamente parecidas tengan **representaciones similares**

# Fundamentos

- Representación de palabras

- Aún siendo una simplificación podemos intuir **dos operaciones importantes en los modelos de lenguaje**

- El mecanismo de **comparación** de palabras (semántico)  
dos palabras son parecidas si tienen muchas coincidencias (productor escalar)
- Podemos **transformar** una palabra para llegar a otra mediante manipulaciones algebraicas (aproximadamente claro):

**rey - hombre + mujer -> (está cerca de) reina**

- Otras analogías obtenidas (word2vec):

<b>tenis - raqueta + bate</b>	<b>-&gt; béisbol</b>
<b>motocicleta - motor + pedales</b>	<b>-&gt; bicicleta</b>
<b>amar - positivo + negativo</b>	<b>-&gt; odiar</b>
<b>parís - francia + españa</b>	<b>-&gt; madrid</b>

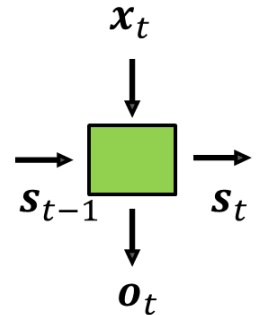
palabra	rey	reina	gato	árbol	ráiz	gustará	odiaba
-----							
propiedades							
persona	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>					
monarquía	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>					
vegetal				<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		
animal			<input checked="" type="checkbox"/>				
-----							
masculino	<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>			
femenino		<input checked="" type="checkbox"/>			<input checked="" type="checkbox"/>		
positivo						<input checked="" type="checkbox"/>	
negativo							<input checked="" type="checkbox"/>
-----							
verbo						<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
presente						<input checked="" type="checkbox"/>	
pasado							<input checked="" type="checkbox"/>
futuro						<input checked="" type="checkbox"/>	

# Fundamentos

- **Redes recurrentes**

- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780

- **Limitación:** *cada celda tiene una dimensión finita para recibir información de los instantes previos y escribir nueva información para el futuro*

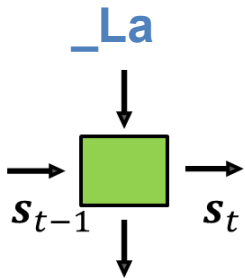
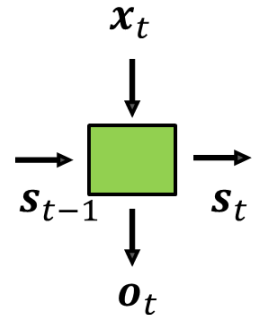


# Fundamentos

- **Redes recurrentes**

- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780

- **Limitación:** cada celda tiene una dimensión finita para recibir información de los instantes previos y escribir nueva información para el futuro



token	prob
unch	(0.21)
uren	(0.09)
wn	(0.03)
ure	(0.03)
...	

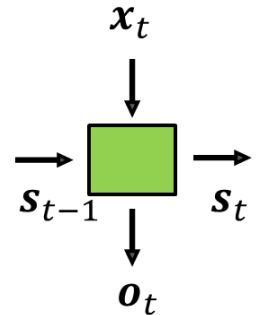


# Fundamentos

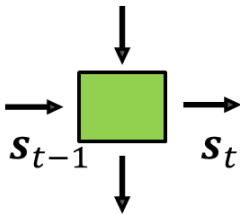
- **Redes recurrentes**

- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780

- **Limitación:** cada celda tiene una dimensión finita para recibir información de los instantes previos y escribir nueva información para el futuro



**\_La \_nueva**



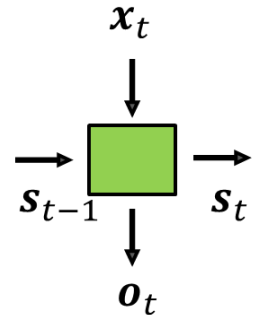
```
token | prob
-----
_versión (0.04)
_gener (0.03)
_est (0.02)
_c (0.01)
...
```

# Fundamentos

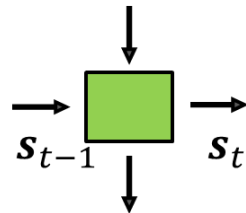
- **Redes recurrentes**

- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780

- **Limitación:** cada celda tiene una dimensión finita para recibir información de los instantes previos y escribir nueva información para el futuro



**\_La** **\_nueva** **\_resol**



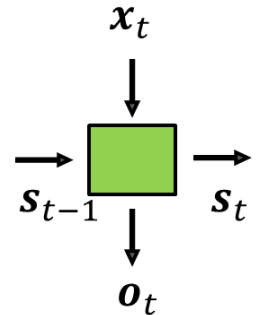
token	prob
ución	(0.96)
ucion	(0.03)
u	(0.01)
uc	(0.00)
...	

# Fundamentos

- **Redes recurrentes**

- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780

- **Limitación:** cada celda tiene una dimensión finita para recibir información de los instantes previos y escribir nueva información para el futuro



2015 <http://karpathy.github.io/2015/05/21/rnn-effectiveness/>

- Red recurrente (LSTM) 3 capas dimensión 512, **5.8M**
- **4.4MB** texto Shakespeare

DUKE VINCENTIO: Well, your wit is in the care of side and that.

Second Lord: They would be ruled after this chamber, and my fair nues begun out of the fact, to be conveyed, Whose noble souls I'll have the heart of the wars.

Clown: Come, sir, I will make did behold your worship.

VIOLA: I'll drink it.

2016 <https://medium.com/deep-writing/harry-potter-written-by-artificial-intelligence-8a9431803da6>

- Primeros 4 libros

Ron didn't even upset her little ingredients on the toilet, and a group of third-year girls last year. Highly bushy and then burst away from them quickly.

"Thought you're all right?" he said.

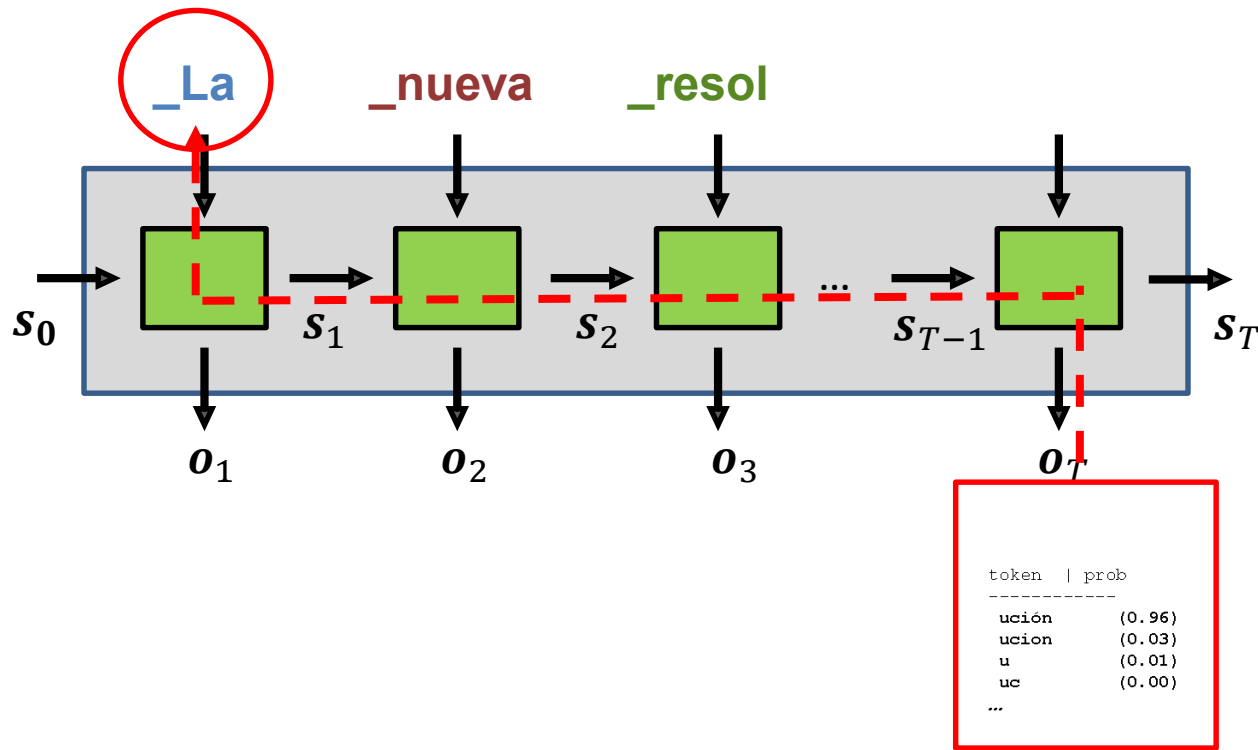
Harry grinned at Harry. "Why should she be cheerful so while you gave detentions, Moody!"

# Fundamentos

- Redes recurrentes

- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780

- **Limitación:** *Gradiente poco robusto para secuencias muy largas*



# Fundamentos

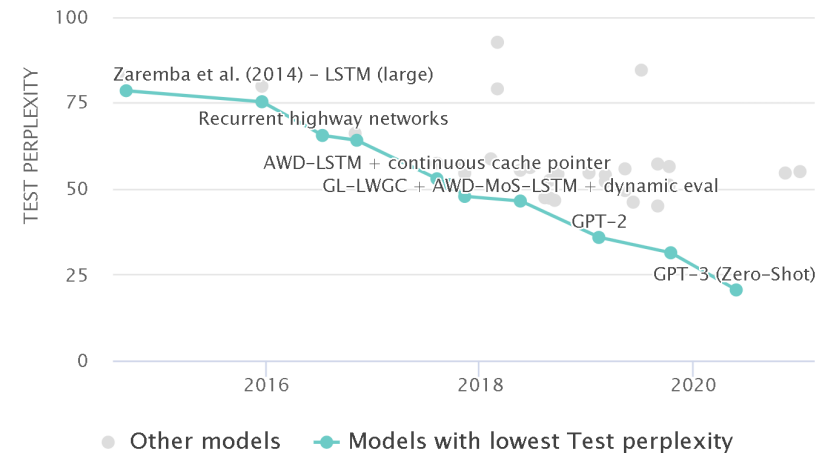
- Transformers
  - Modelo presentado por Google en 2017
  - “Attention is all you need” 78k citas

---

## Attention Is All You Need

---

<b>Ashish Vaswani*</b> Google Brain avaswani@google.com	<b>Noam Shazeer*</b> Google Brain noam@google.com	<b>Niki Parmar*</b> Google Research nikip@google.com	<b>Jakob Uszkoreit*</b> Google Research usz@google.com
<b>Llion Jones*</b> Google Research llion@google.com	<b>Aidan N. Gomez* †</b> University of Toronto aidan@cs.toronto.edu	<b>Lukasz Kaiser*</b> Google Brain lukaszkaier@google.com	
<b>Illia Polosukhin* †</b> illia.polosukhin@gmail.com			

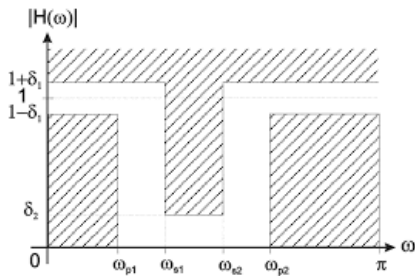


- ¿Por qué ha tenido tanto éxito?



# Fundamentos

- Transformers: ¿Máquinas de propósito general?
  - Planteamiento clásico de un problema de ingeniería

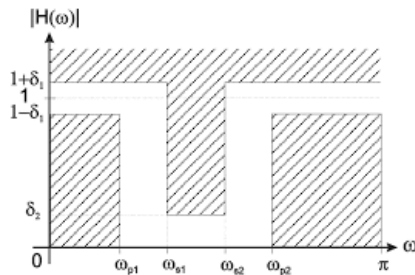


- se plantean unos **requisitos**/restricciones
- se elige un **método** para la solución, a veces de forma analítica, otras aproximada, otras por optimización iterativa...
- se **comprueba** si la solución cumple con las restricciones

# Fundamentos

- Transformers: ¿Máquinas de propósito general?

- Planteamiento clásico de un problema de ingeniería



- se plantean unos requisitos/restricciones
- se elige un método para la solución, a veces de forma analítica, otras aproximada, otras por optimización iterativa...
- se comprueba si la solución cumple con las restricciones

- En aprendizaje automático presentamos datos y la solución esperada

- Por ejemplo dar varios números aleatorios y como objetivo querer los números ordenados

- **Preparamos muchos ejemplos** distintos y entrenamos un sistema

- Medimos la calidad, **no siempre acierta el 100%** de los casos!

entrada	salida
0, 2, 1, 5	0, 1, 2, 5
4, 5, 3	3, 4, 5
0, 1	0, 1
8, 4, 5	4, 5, 8
...	

# Fundamentos

- Transformers: ¿Máquinas programables de propósito general?
  - “differentiable computer”

- 2014, Neural Turing Machines
- 2016, Differentiable Neural Computer

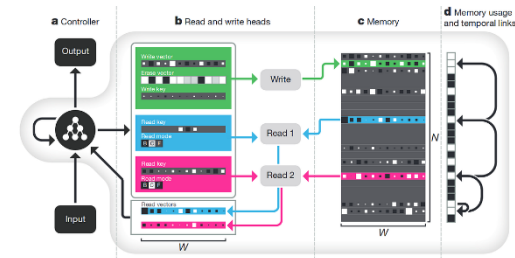
## nature

### Hybrid computing using a neural network with dynamic external memory

[Alex Graves](#), [Greg Wayne](#), [Malcolm Reynolds](#), [Tim Harley](#), [Ivo Danihelka](#), [Agnieszka Grabska-Barwińska](#), [Sergio Gómez Colmenarejo](#), [Edward Grefenstette](#), [Tiago Ramalho](#), [John Agapiou](#), [Adrià Puigdomènech Badia](#), [Karl Moritz Hermann](#), [Yori Zwols](#), [Georg Ostrovski](#), [Adam Cain](#), [Helen King](#), [Christopher Summerfield](#), [Phil Blunsom](#), [Koray Kavukcuoglu](#) & [Demis Hassabis](#)

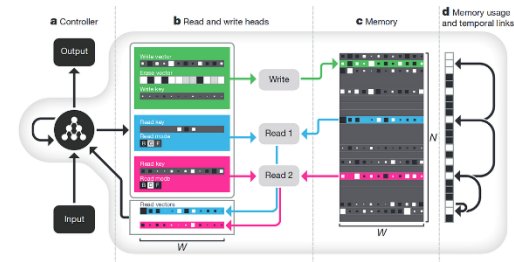
- Tenía memorías diferenciables:
  - » Escritura / Lectura
  - » **Heads / cabezales**

- 2017, Transformers

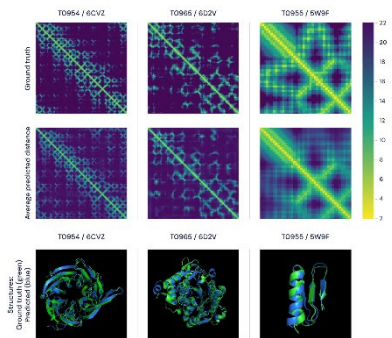
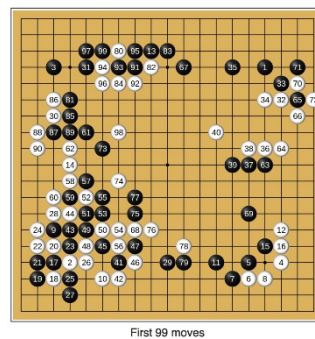
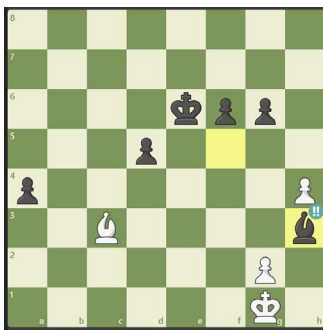


# Fundamentos

- Transformers: ¿Máquinas programables de propósito general?
  - “differentiable computer”
    - 2014, Neural Turing Machines
    - 2016, Differentiable Neural Computer
    - 2017, Transformers



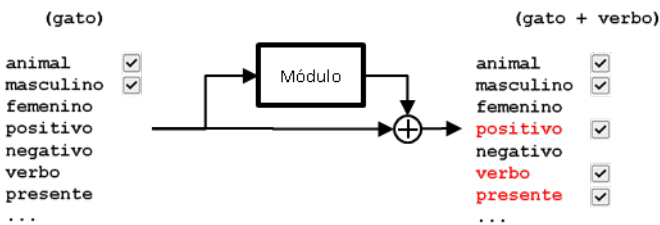
- ¿Qué utilidad tiene aprenda un algoritmo que ya existe? (y que no siempre lo haga bien)
  - Estos experimentos se han hecho para demostrar esa capacidad de generar pequeños algoritmos si se necesitan
  - Esta capacidad destaca cuando **no se conoce un algoritmo** para resolver la tarea: traducir entre idiomas, resumir textos, evaluar una posición estratégica, predecir el plegado de proteínas...



<https://www.deepmind.com/blog/alphafold-using-ai-for-scientific-discovery-2020>

# Fundamentos

- Transformers: 3 ideas

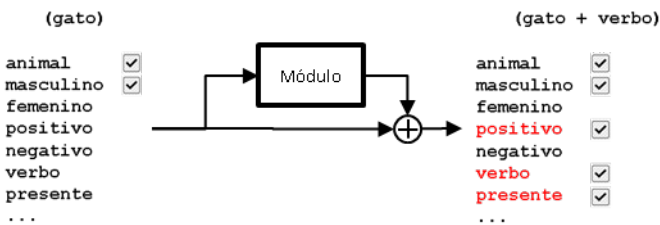


La información se procesa  
**incrementalmente** (residual)

*Thought vectors (G. Hinton)*

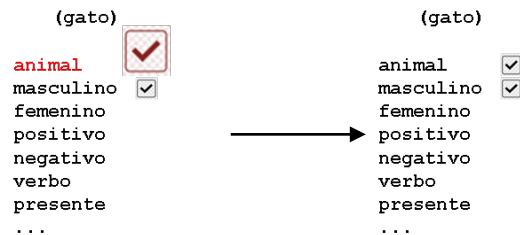
# Fundamentos

- Transformers: 3 ideas



La información se procesa **incrementalmente** (residual)

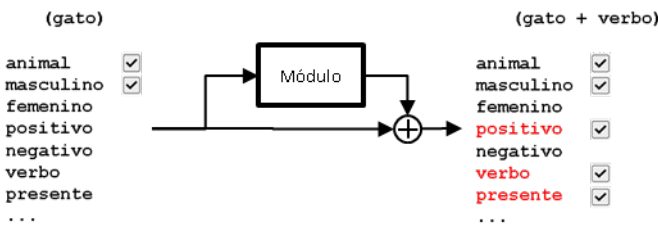
*Thought vectors (G. Hinton)*



Es crítico para que funcione **normalizar la escala** de los vectores

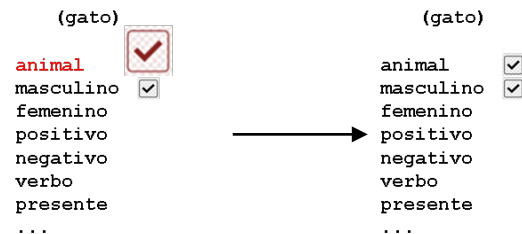
# Fundamentos

- Transformers: 3 ideas

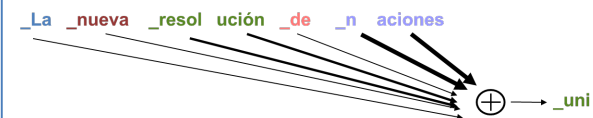


La información se procesa **incrementalmente** (residual)

*Thought vectors (G. Hinton)*



Es crítico para que funcione **normalizar la escala** de los vectores



Dos tipos de módulos  
Procesado de vectores individuales

- red neuronal

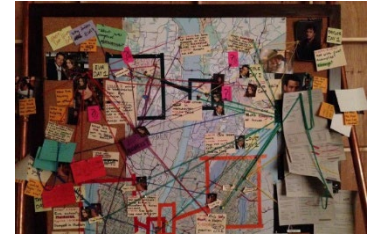
Mezcla de vectores

- mecanismo de atención** (auto)  
(suma ponderada)



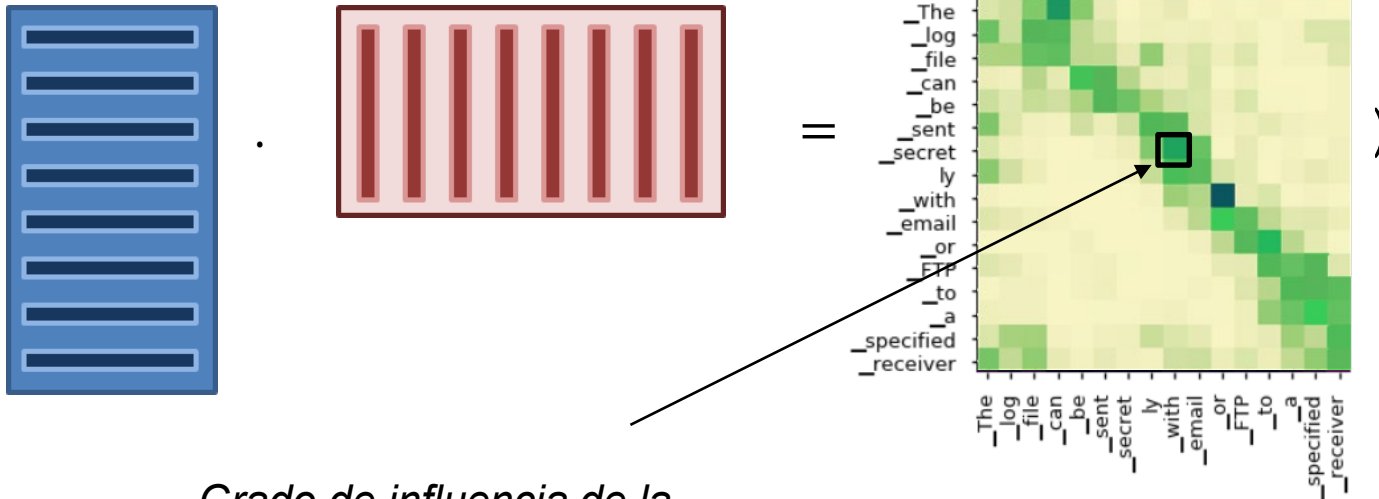
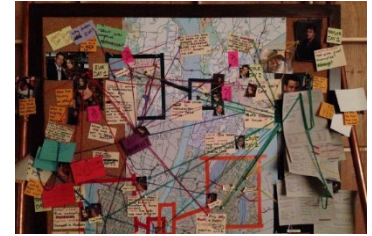
# Fundamentos

- Mecanismo de atención (auto)
  - Son capaces de analizar la relación de todas las palabras
  - ¿Cómo se genera esas relaciones?



# Fundamentos

- Mecanismo de atención (auto)
  - Son capaces de analizar la relación de todas las palabras
  - ¿Cómo se genera esas relaciones?



*Grado de influencia de la entrada  $i$  sobre la  $j$*

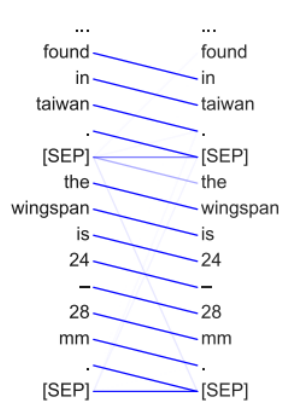
<https://nlp.seas.harvard.edu/2018/04/03/attention.html>

RNN

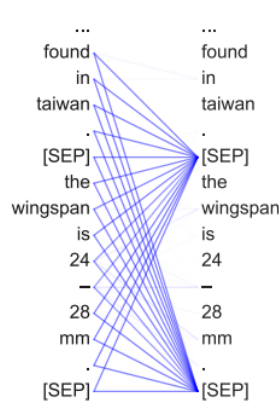
# Fundamentos

- Transformers

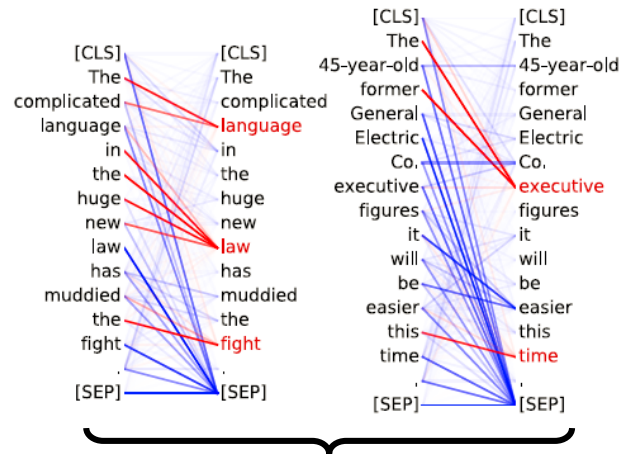
- Son capaces de analizar la relación de todas las entradas, múltiples aspectos: heads (cabezales)
  - por ejemplo en LLAMA 7b -> 32 (en 32 capas)



palabra anterior



Final de frase



Determinantes y modificadores de un nombre



Desambigüación ellos -> vecinos

<https://medium.com/dair-ai/aspects-of-language-captured-by-bert-32bc3c54016f>

<https://medium.com/synapse-dev/understanding-bert-transformer-attention-isnt-all-you-need-5839ebd396db>

RNN

VIVOLAB



Universidad Zaragoza

# Fundamentos

Datos GPT-3 13B / LLAMA13B

\_La \_nueva \_resol ución \_de \_n aciones



**Transformer  
(decoder)**

	1	2	3	4	5	6	7	8	9	10
1	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0

1. Extraer vector de la tabla/diccionario

# Fundamentos

Datos GPT-3 13B / LLAMA13B

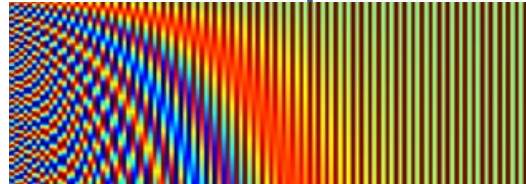
\_La \_nueva \_resol ución \_de \_n aciones

**Transformer  
(decoder)**

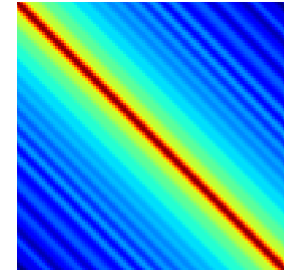





1.2 Sumar marcadores de posición



(distancia relativa)



# Fundamentos

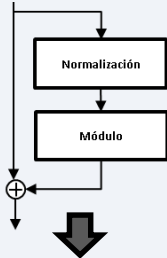
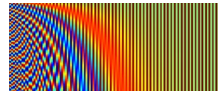
Datos GPT-3 13B / LLAMA13B

\_La \_nueva \_resol ución \_de \_n aciones

**Transformer  
(decoder)**



...	...	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...



- 2.1. Normalización
- 2.2. Procesado intra-token

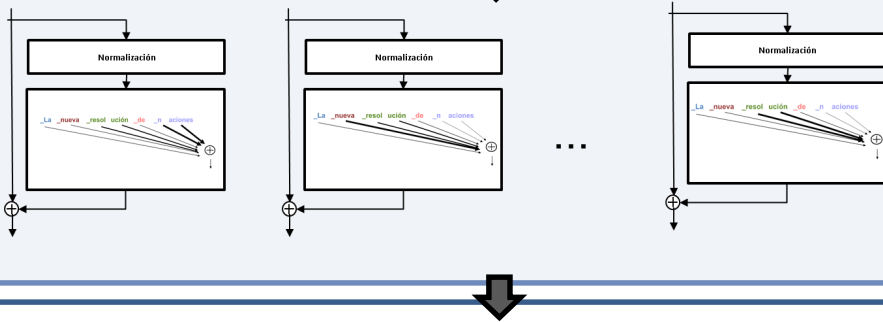
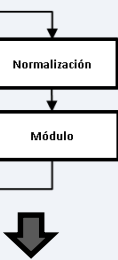
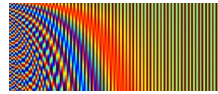
# Fundamentos

Datos GPT-3 13B / LLAMA13B

\_La \_nueva \_resol ución \_de \_n aciones

Transformer  
(decoder)



x 40 (heads / cabezales)

- 3.1 Normalización Self Attention
- 3.2 Procesado inter token

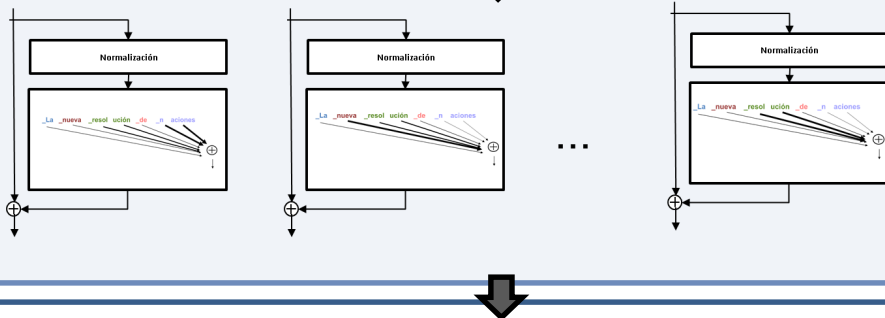
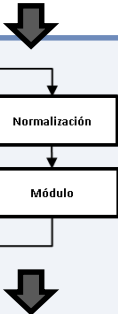
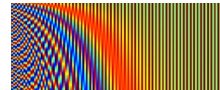


# Fundamentos

Datos GPT-3 13B / LLAMA13B

\_La \_nueva \_resol ución \_de \_n aciones

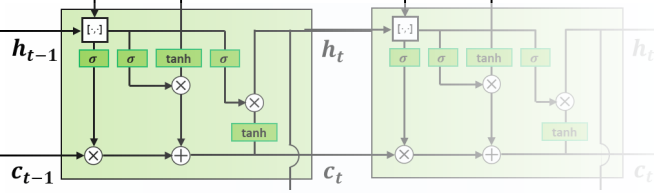
**Transformer  
(decoder)**



x 40 capas

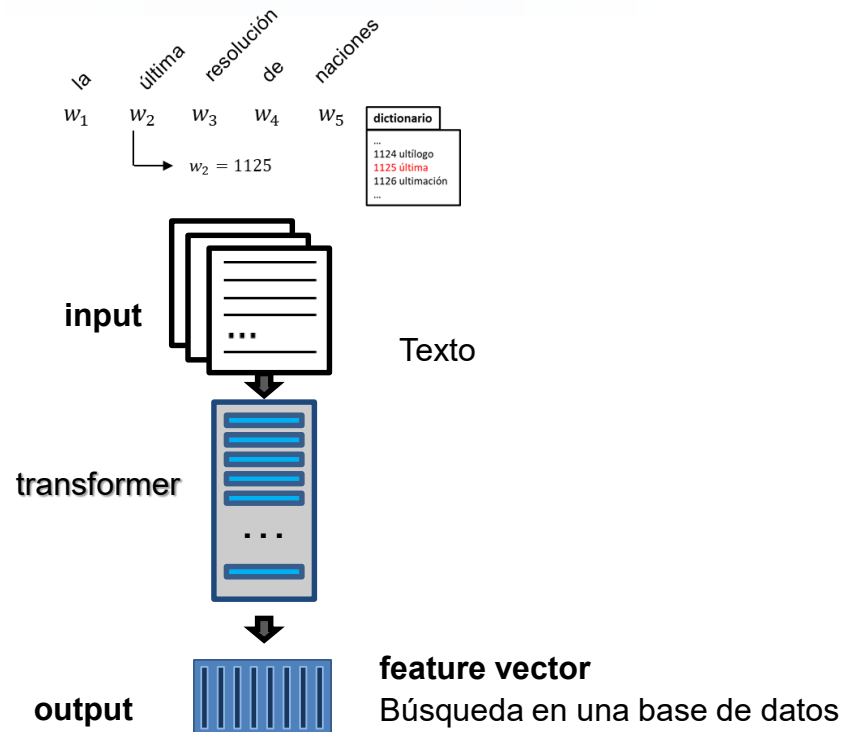
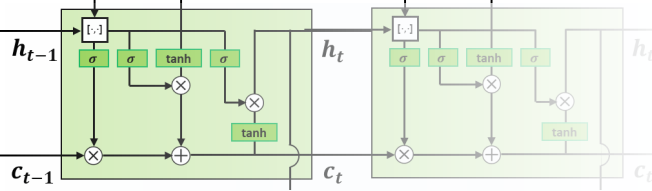
x 40 (heads / cabezales)

# Aplicaciones



La versatilidad de los transformer:  
¿ Qué tareas pueden resolver ?

# Aplicaciones



## – Detección de autoría/fraude

- Arquitectura similar a los de extracción de representación en biometría se trata de decir si un documento o transcripción de reconocedor implica fraude o si dos textos los ha producidos la misma persona.

## – Reconocimiento de voz:

- Transcripción de audios a texto.

## – Generación de texto/Generación de voz

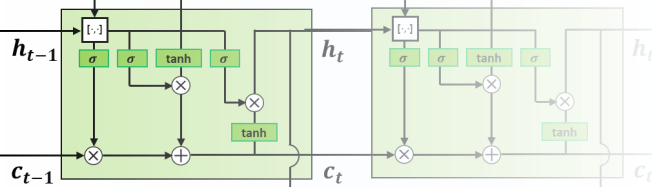
- Modelo entrenado con el objetivo de **predecir la siguiente palabra** o siguiente muestra

## – Biometría de voz:

- Reconocer a un locutor por la voz en sistemas de acceso con contraseña vocal

## – Búsqueda basada en vectores

- Un modelo a gran escala de texto tiene la capacidad de representar en unos pocos valores el contexto de la información
- Se puede utilizar esta **descripción no supervisada** para buscar situaciones parecidas en un repositorio



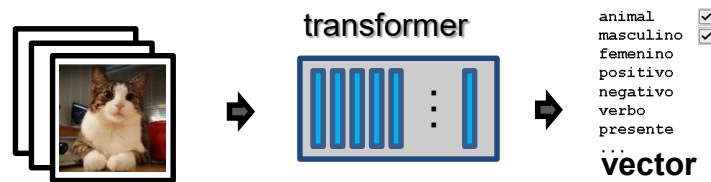
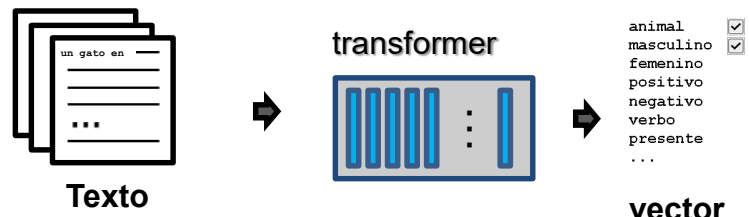
# Aplicaciones

- En 2021 OpenAi presentó: CLIP

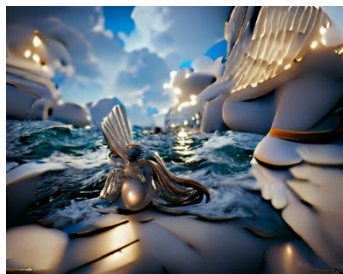
## Learning Transferable Visual Models From Natural Language Supervision

Alec Radford<sup>\*1</sup> Jong Wook Kim<sup>\*1</sup> Chris Hallacy<sup>1</sup> Aditya Ramesh<sup>1</sup> Gabriel Goh<sup>1</sup> Sandhini Agarwal<sup>1</sup>  
 Girish Sastry<sup>1</sup> Amanda Askell<sup>1</sup> Pamela Mishkin<sup>1</sup> Jack Clark<sup>1</sup> Gretchen Krueger<sup>1</sup> Ilya Sutskever<sup>1</sup>

- imagen y texto:
  - **espacio de representación común**



**vectores similares**



the angel of air. unreal engine  
[@arankomatsuzaki](#)



treehouse in the style of studio Ghibli  
 animation [@danielrussrus](#)

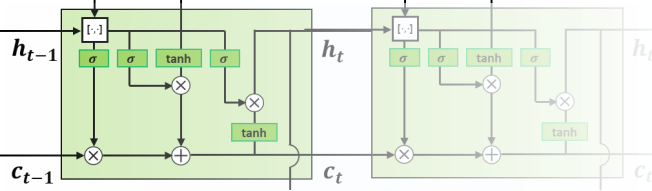


A wooden spanish laptop of 1650  
 found the library of El Escorial



Medieval 1230 book page illustrating  
 monks playing basketball

# Aplicaciones



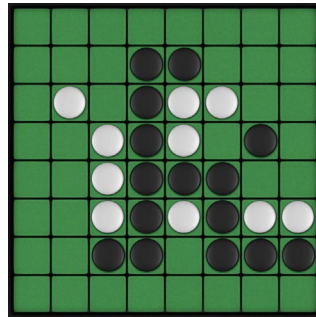
- **Ejemplo** generación de modelos del mundo

- El modelo se entrena para predecir en siguiente movimiento.

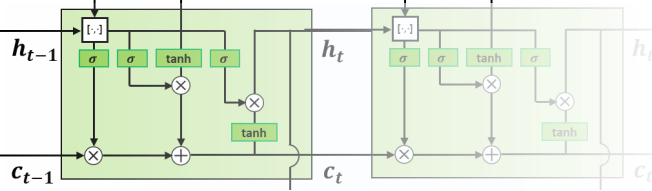
- El modelo tiene una secuencia de movimientos

F4, F3, D2, F5, G2, F2, G3, C4, E5, F6, D6, E2, B4, C5, G7, C1, G6, F7, G5, C3, B3, **H6**

- Si nos ayudamos de una representación visual entendemos mucho mejor de qué juego se trata y cuál es el estado actual del tablero



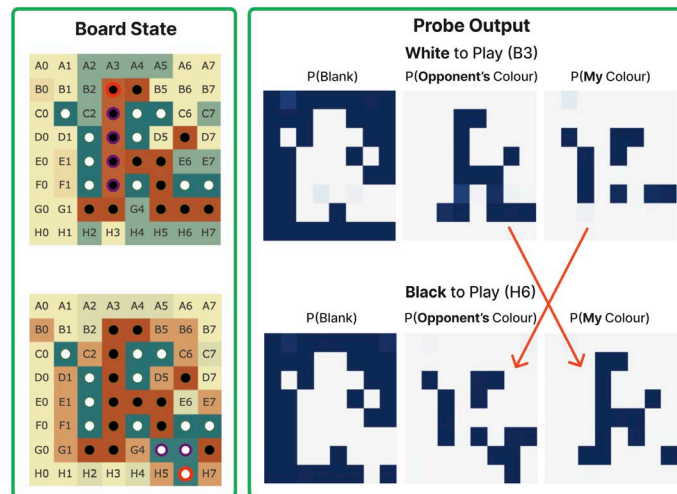
# Aplicaciones



- **Ejemplo** generación de modelos del mundo

- En ese ejemplo se muestra cómo procesando los vectores de representación.

- El modelo representa internamente el mapa del tablero diferenciando el color de las piezas y a quién le toca jugar



<https://www.lesswrong.com/posts/nmxzr2zsjNtjaHh7x/actually-othello-gpt-has-a-linear-emergent-world>

# Aplicaciones

- Transformers: Razonamiento / modelos del mundo
  - Existen numerosos trabajos recientes en los que se evalúan o entrenan LMs para resolver multitud de tareas

### Task 1: Single Supporting Fact

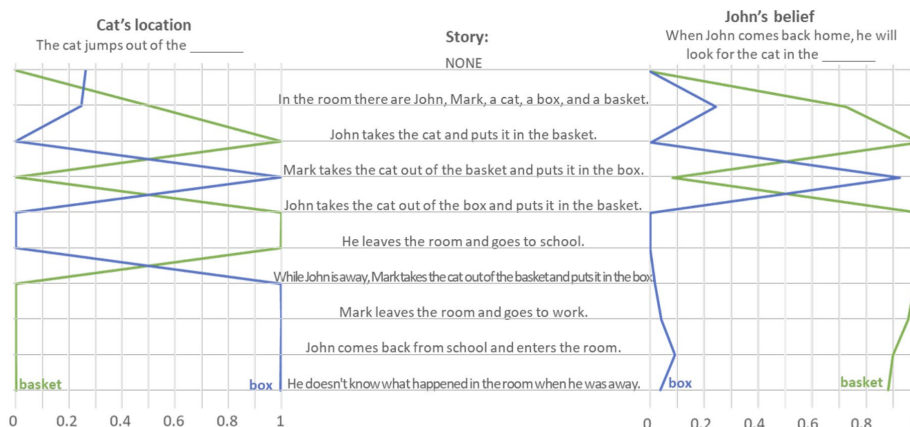
Mary went to the bathroom.  
John moved to the hallway.  
Mary travelled to the office.  
Where is Mary? A:office

### Task 7: Time Reasoning

In the afternoon Julie went to the park.  
Yesterday Julie was at school.  
Julie went to the cinema this evening.  
Where did Julie go after the park? A:cinema  
Where was Julie before the park? A:school

### Task 8: Positional Reasoning

The triangle is to the right of the blue square.  
The red square is on top of the blue square.  
The red sphere is to the right of the blue square.  
Is the red sphere to the right of the blue square? A:yes  
Is the red square to the left of the triangle? A:yes



Xiang, J et al (2023). Language Models Meet World Models: Embodied Experiences Enhance Language Models.

Theory of Mind May Have Spontaneously Emerged in Large Language Models Authors: Michal Kosinski\*1





# Modelos de lenguaje

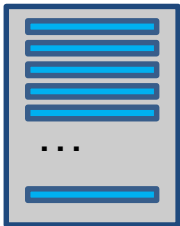
- Modelos de lenguaje, chats y agentes

# Modelos de lenguaje

- **Entrenamiento**

- Llamamos entrenar/optimizar/ajustar/aprender

\_La \_nueva \_resol ución \_de \_n acciones...



transformer

Como modificamos cada parámetro para que

- aumente la probabilidad ?
- baje el coste?

\_un

aumentar probabilidad



el coste es negativo de la probabilidad ?



$p(\text{\_un} | \text{\_La}, \text{\_nueva}, \text{\_resol}, \text{\_ución}, \text{\_de}, \text{\_n}, \text{\_aciones})$

Proceso muy costoso en el caso de los LLMs

**miles de millones de parámetros**



# Modelos de lenguaje

- **Entrenamiento (no supervisado)**

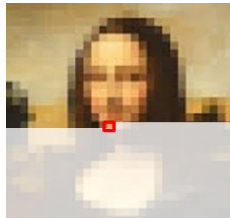
- **Coste computacional alto**

- Entrenar un LLM puede suponer usar cientos o miles de GPUs durante semanas (costes de millones de \$)

- **No supervisado**

- Lo que lo hace viable es que no se necesitan etiquetas
- Gran cantidad de textos recopilados de distintas fuentes: Wikipedia, Periódicos, Web, Github

- **También usado en imagen, audio ...**



Viendo los pixels anteriores:  
¿cómo es el siguiente ?

- **Entrenamiento (supervisado)**

- **Disponer: preguntas y sus respuestas/etiquetas**

- Alto coste en etiquetado
- Adaptación a comportarse como un chabot

- **Problema sesgos en los datos**

- Si mostramos más veces un ejemplo y la respuesta que otros ejemplos aparecerá un sesgo en el sistema

- **Etiquetado:**



*Todos nosotros etiquetamos / Instalaciones especializadas*

# Modelos de lenguaje

- Chats y Agentes derivados de LLMs

'Si tenemos dos manzanas y tres peras, en total tenemos '



'... la posibilidad de cocinar ...'

LLM (foundation model)

- gpt2,3,4
- bloom
- llama
- falcon
- ...

prompt

modelo

salida

# Modelos de lenguaje

- Chats y Agentes derivados de LLMs

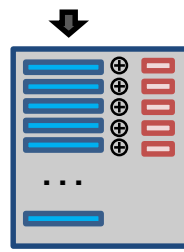
'Si tenemos dos manzanas y tres peras, en total tenemos '



... la posibilidad de cocinar ...'

LLM (foundation model)

- gpt2,3,4
- bloom
- llama
- falcon
- ...



La respuesta es cinco frutas.'

Chat – asistente

- Instruct T5
- chatgpt
- alpaca
- vicuna
- ...

modelo base

- T5
- gpt
- llama
- llama

Adaptación:

- finetuning
- LORA + cuantificación
- Reinforcement Learning

Frases asistente (variedad)

Lo siento pero ...

Por supuesto aquí tienes ...

Tienes razón ...

prompt

modelo

salida

# Modelos de lenguaje

- Chats y Agentes derivados de LLMs

'Si tenemos dos manzanas y tres peras, en total tenemos '

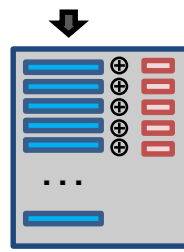


... la posibilidad de cocinar ...'

### LLM (foundation model)

- gpt2,3,4
- bloom
- llama
- falcon
- ...

- Gran coste > varios millones €
- >1000 GPUs
- Semanas de entrenamiento
- Textos de entrenamiento > 1TB
- **entrenamiento no supervisado**
- contenido no verificado



La respuesta es cinco frutas.'

### Chat – asistente

- Instruct T5
- chatgpt
- alpaca
- vicuna
- ...

### modelo base

- T5
- gpt
- llama
- llama

- Coste moderado < 1000 €
- 1-10 GPUs
- Días/horas de entrenamiento
- Textos de entrenamiento < 1GB
- **entrenamiento supervisado**
- contenido verificado

Adaptación:

- finetuning
- LORA + cuantificación
- Rejforcement Learning

Frases asistente (variedad)

Lo siento pero ...

Por supuesto aquí tienes ...

Tienes razón ...

prompt

modelo

salida

# Modelos de lenguaje

- Tamaño del modelo

	año	dime	capas	heads	parámetros (M)	
<hr/> <b>Improving Language Understanding by Generative Pre-Training</b> <hr/> <small>Alec Radford OpenAI alec@openai.com</small> <small>Karthik Narasimhan OpenAI karthika@openai.com</small> <small>Tim Salimans OpenAI tins@openai.com</small> <small>Ilya Sutskever OpenAI ilyasu@openai.com</small>	GPT1	2018	768	12	12	117 M
<hr/> <b>Language Models are Unsupervised Multitask Learners</b> <hr/> <small>Alec Radford<sup>1</sup></small> <small>Jeffrey Wu<sup>1</sup></small> <small>Rewon Child<sup>1</sup></small> <small>David Luan<sup>1</sup></small> <small>Dario Amodei<sup>1</sup></small> <small>Ilya Sutskever<sup>1</sup></small>	GPT2	2019	1600	48	25	1500 M
<hr/> <b>Language Models are Few-Shot Learners</b> <hr/> <small>Tom B. Brown<sup>*</sup></small> <small>Benjamin Mann<sup>*</sup></small> <small>Nick Ryder<sup>*</sup></small> <small>Melanie Subbiah<sup>*</sup></small> <small>Jared Kaplan<sup>1</sup></small> <small>Prafulla Dhariwal<sup>1</sup></small> <small>Arvind Neelakantan<sup>1</sup></small> <small>Pranav Shyam<sup>1</sup></small> <small>Girish Sastry<sup>1</sup></small> <small>Amanda Askell<sup>1</sup></small> <small>Sandhini Agarwal<sup>1</sup></small> <small>Ariel Herbert-Voss<sup>1</sup></small> <small>Gretchen Krueger<sup>1</sup></small> <small>Tom Henighan<sup>1</sup></small> <small>Rewon Child<sup>1</sup></small> <small>Aditya Ramesh<sup>1</sup></small> <small>Daniel M. Ziegler<sup>1</sup></small> <small>Jeffrey Wu<sup>1</sup></small> <small>Clemens Winter<sup>1</sup></small> <small>Christopher Hesse<sup>1</sup></small> <small>Mark Chen<sup>1</sup></small> <small>Eric Sigler<sup>1</sup></small> <small>Mateusz Litwin<sup>1</sup></small> <small>Scott Gray<sup>1</sup></small> <small>Benjamin Chess<sup>1</sup></small> <small>Jack Clark<sup>1</sup></small> <small>Christopher Berner<sup>1</sup></small> <small>Sam McCandlish<sup>1</sup></small> <small>Alec Radford<sup>1</sup></small> <small>Ilya Sutskever<sup>1</sup></small> <small>Dario Amodei<sup>1</sup></small>	GPT3	2020	4096	32	32	6700 M (llama7b)
			5140	40	40	13000 M (llama13b)
			12288	96	96	175000 M



# Modelos de lenguaje

- Tamaño del modelo

	año	dime	capas	heads	parámetros (M)	
<hr/> <b>Improving Language Understanding by Generative Pre-Training</b> <hr/> <small>Alec Radford OpenAI alec@openai.com</small> <small>Karthik Narasimhan OpenAI karthika@openai.com</small> <small>Tim Salimans OpenAI tins@openai.com</small> <small>Ilya Sutskever OpenAI ilyasu@openai.com</small>	GPT1	2018	768	12	12	117 M
<hr/> <b>Language Models are Unsupervised Multitask Learners</b> <hr/> <small>Alec Radford<sup>1</sup></small> <small>Jeffrey Wu<sup>1</sup></small> <small>Rewon Child<sup>1</sup></small> <small>David Luan<sup>1</sup></small> <small>Dario Amodei<sup>1</sup></small> <small>Ilya Sutskever<sup>1</sup></small>	GPT2	2019	1600	48	25	1500 M
<div style="border: 1px solid black; padding: 5px; background-color: #f0f0f0;"> <b>OpenAI decidió no liberar los modelos de GPT2 de mayor tamaño por miedo a los posibles malos usos (2019)</b> </div>						
<hr/> <b>Language Models are Few-Shot Learners</b> <hr/> <small>Tom B. Brown<sup>*</sup></small> <small>Benjamin Mann<sup>*</sup></small> <small>Nick Ryder<sup>*</sup></small> <small>Melanie Subbiah<sup>*</sup></small> <small>Jared Kaplan<sup>1</sup></small> <small>Prafulla Dhariwal<sup>1</sup></small> <small>Arvind Neelakantan<sup>1</sup></small> <small>Pranav Shyam<sup>1</sup></small> <small>Girish Sastry<sup>1</sup></small> <small>Amanda Askell<sup>1</sup></small> <small>Sandhini Agarwal<sup>1</sup></small> <small>Ariel Herbert-Voss<sup>1</sup></small> <small>Gretchen Krueger<sup>1</sup></small> <small>Tom Henighan<sup>1</sup></small> <small>Rewon Child<sup>1</sup></small> <small>Aditya Ramesh<sup>1</sup></small> <small>Daniel M. Ziegler<sup>1</sup></small> <small>Jeffrey Wu<sup>1</sup></small> <small>Clemens Winter<sup>1</sup></small> <small>Christopher Hesse<sup>1</sup></small> <small>Mark Chen<sup>1</sup></small> <small>Eric Sigler<sup>1</sup></small> <small>Matusz Litwin<sup>1</sup></small> <small>Scott Gray<sup>1</sup></small> <small>Benjamin Chess<sup>1</sup></small> <small>Jack Clark<sup>1</sup></small> <small>Christopher Berner<sup>1</sup></small> <small>Sam McCandlish<sup>1</sup></small> <small>Alec Radford<sup>1</sup></small> <small>Ilya Sutskever<sup>1</sup></small> <small>Dario Amodei<sup>1</sup></small>	GPT3	2020	4096	32	32	6700 M (llama7b)
			5140	40	40	13000 M (llama13b)
			12288	96	96	175000 M

# Modelos de lenguaje

- **GPT2: Language Models are Unsupervised Multitask Learners**
  - La gran capacidad de almacenamiento de datos concretos
  - La coherencia de los textos generados incluso alejados del entrenamiento

**Context (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.

**GPT-2:** 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.

prompt

generación

# Modelos de lenguaje

- GPT3: Language Models are Few-Shot Learners
  - Seguir instrucciones del propio prompt para definir la tarea
  - La capacidad de aprender de ejemplos del prompt: few-shot (meta learning)

Context →	Final Exam with Answer Key Instructions: Please carefully read the following passages. For each passage, you must identify which noun the pronoun marked in <b>*bold*</b> refers to. ===== 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 <b>*his*</b> financial support. Question: In the passage above, what does the pronoun " <b>*his*</b> " refer to? Answer:	} prompt
Target Completion →	mr. moncrieff	
Poor English input:	I eated the purple berries.	} prompt
Good English output:	I ate the purple berries.	
Poor English input:	Thank you for picking me as your designer. I'd appreciate it.	} generación
Good English output:	Thank you for choosing me as your designer. I appreciate it.	
Poor English input:	The mentioned changes have done. or I did the alteration that you requested. or I changed things you wanted and did the modifications.	} prompt
Good English output:	The requested changes have been made. or I made the alteration that you requested. or I changed things you wanted and made the modifications.	
Poor English input:	I'd be more than happy to work with you in another project.	} generación
Good English output:	I'd be more than happy to work with you on another project.	

# Modelos de lenguaje

- Modelos de chat, GPT4, ORCA13B
  - Aprueban exámenes de todo tipo
  - Resultados más altos en tareas de razonamiento, sentido común, causalidad

Task	ChatGPT	GPT-4	Vicuna-13B	Orca-13B
Boolean Expressions	82.8	77.6	40.8	72.0 (76.5%)
Causal Judgement	57.2	59.9	42.2	59.9 (41.8%)
Date Understanding	42.8	74.8	10.0	50.0 (400.0%)
Disambiguation QA	57.2	69.2	18.4	63.6 (245.7%)
Formal Fallacies	53.6	64.4	47.2	56.0 (18.6%)
Geometric Shapes	25.6	40.8	3.6	20.8 (477.8%)
Hyperbaton	69.2	62.8	44.0	64.0 (45.5%)
Logical Deduction (5 objects)	38.8	66.8	4.8	39.6 (725.0%)
Logical Deduction (7 objects)	39.6	66.0	1.2	36.0 (2900.0%)
Logical Deduction (3 objects)	60.4	94.0	16.8	57.6 (242.9%)
Movie Recommendation	55.4	79.5	43.4	78.3 (80.6%)

### Orca: ✓

Assistant: Let the lengths of the two legs be  $5x$  and  $12x$ . Since the hypotenuse is the sum of the squares of the lengths of the two legs, we can write the equation:

$$(5x)^2 + (12x)^2 = 13^2$$

First, square the lengths of the legs:

$$25x^2 + 144x^2 = 169$$

Combine the terms:

$$169x^2 = 169$$

Now, divide by 169:

$$x^2 = 1$$

Since  $x^2$  must be a whole number,  $x$  can only be 1 or -1. However, since the lengths of the legs must be positive, we can discard the negative value of  $x$ . So, the lengths of the two legs are 5 and 12.

### System:

You are an AI assistant that follows instruction extremely well. Help as much as you can.

### Human:

Given that the hypotenuse of a right triangle is 13, and the ratio of the lengths of the two legs is 5:12, find the lengths of the two legs.

### GPT-4: ✓

Let's denote the length of the shorter leg as  $a$  and the length of the longer leg as  $b$ . According to the given information, we have the following relationships:

1.  $a^2 + b^2 = 13^2$  (Pythagorean theorem)

2.  $\frac{a}{b} = \frac{5}{12}$  (given ratio)

Now, we can solve for  $a$  and  $b$  using these relationships.

First, let's solve the second equation for  $a$ :

$$a = \frac{5}{12} \cdot b$$

Now, substitute this expression for  $a$  into the first equation:

$$\left(\frac{5}{12} \cdot b\right)^2 + b^2 = 13^2$$

Simplify the equation:

$$\frac{25}{144} \cdot b^2 + b^2 = 169$$

Now, find a common denominator for the fractions:

$$\frac{25b^2 + 144b^2}{144} = 169$$

*Mukherjee, S., et al (2023). Orca: Progressive Learning from Complex Explanation Traces of GPT-4. arXiv preprint arXiv:2306.02707.*

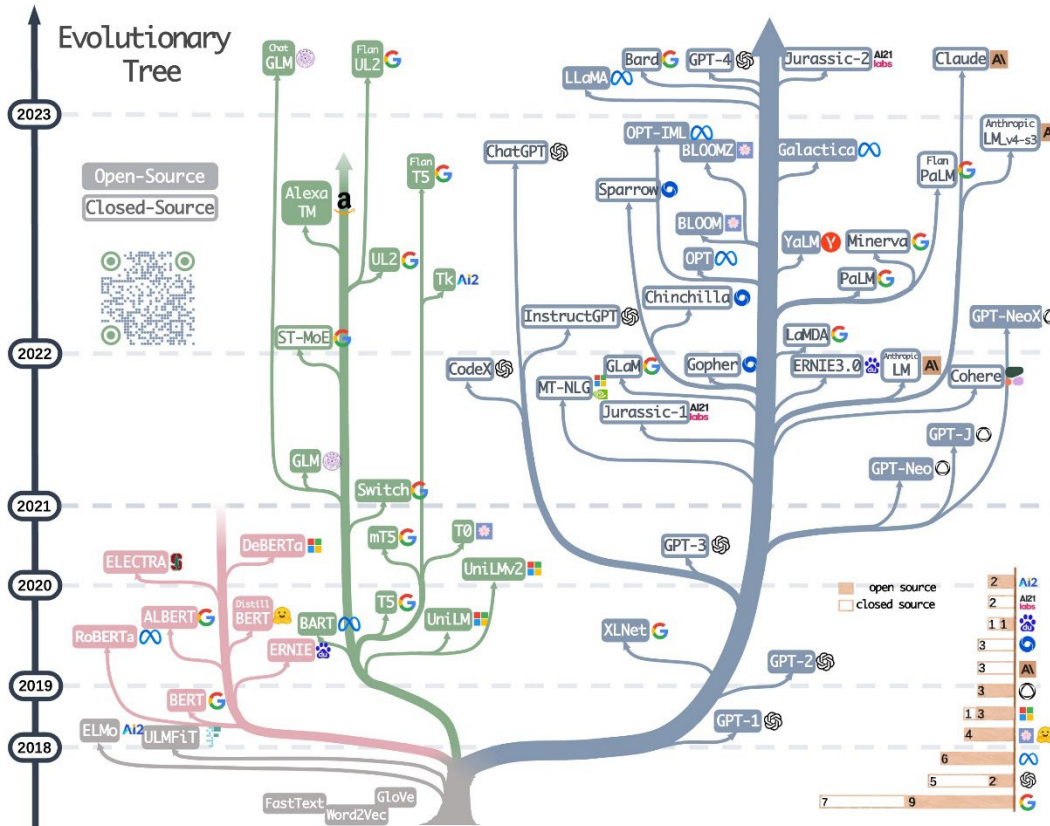
# Modelos de lenguaje

- Modelos de chat, Alpaca, Vicuna ....
  - Muy fáciles de entrenar, pocos recursos
  - **Ejemplo Entrenamiento**

```
{
  "instruction": "Give three tips for staying healthy.",
  "input": "",
  "output": "1. Eat a balanced diet and make sure to include plenty of fruits and
vegetables. \n2. Exercise regularly to keep your body active and strong. \n3. Get enough
sleep and maintain a consistent sleep schedule."
},
{
  "instruction": "What are the three primary colors?",
  "input": "",
  "output": "The three primary colors are red, blue, and yellow."
},
...
{
  "instruction": "Find the derivative of  $f(x) = x^2 - 4x + 6$ .",
  "input": "",
  "output": "The derivative of  $f(x) = x^2 - 4x + 6$  is  $f'(x) = 2x - 4$ ."
},
},
```

# Modelos de lenguaje

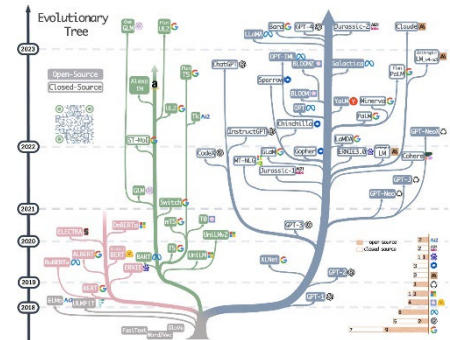
- Ecosistema LLMs: explosión



<https://github.com/Mooler0410/LLMsPracticalGuide>

# Modelos de lenguaje

- Ecosistema LLMs: explosión
- Y la lista no para de crecer
  - ORCA13b (de LLAMA13b)
  - Falcon 7b, 40b (UAE)
  - China 79 LLMs
    - 2020: 2 LLMs
    - 2021: 30 LLMs
    - 2022: 28 LLMs
    - 2023: 19 LLMs (hasta junio)
- ...
- phi-1 (1.3b) Textbooks Are All You Need (20 Jun 2023)





# Modelos de lenguaje

- Ecosistema LLMs: explosión

**A12 Reasoning Challenge (25-shot)** grade-school science questions.  
**HellaSwag (10-shot)** commonsense inference, easy for humans (~95%)  
**MMLU (5-shot)** elementary mathematics, US history, computer science, law  
**TruthfulQA (0-shot)** measure propensity to reproduce falsehood

Model	Revision	Average	ARC (25-s)	HellaSwag (10-s)	MMLU (5-s)	TruthfulQA (MC) (0-s)
<a href="#">tiiuae/falcon-40b-instruct</a>	main	63.2	61.6	84.4	54.1	52.5
<a href="#">timdettmers/guanaco-65b-merged</a>	main	62.2	60.2	84.6	52.7	51.3
<a href="#">CalderaAI/30B-Lazarus</a>	main	60.7	57.6	81.7	45.2	58.3
<a href="#">tiiuae/falcon-40b</a>	main	60.4	61.9	85.3	52.7	41.7
<a href="#">timdettmers/guanaco-33b-merged</a>	main	60	58.2	83.5	48.5	50
<a href="#">ausboss/llama-30b-supercot</a>	main	59.8	58.5	82.9	44.3	53.6
<a href="#">huggyllama/llama-65b</a>	main	58.3	57.8	84.2	48.8	42.3
<a href="#">pinkmanlove/llama-65b-hf</a>	main	58.3	57.8	84.2	48.8	42.3
<a href="#">llama-65b</a>	main	58.3	57.8	84.2	48.8	42.3
<a href="#">MetaIX/GPT4-X-Alpaca-30b</a>	main	57.9	56.7	81.4	43.6	49.7
<a href="#">Aeala/VicUnlocked-alpaca-30b</a>	main	57.6	55	80.8	44	50.4
<a href="#">digitous/Alpaca30b</a>	main	57.4	57.1	82.6	46.1	43.8
<a href="#">Aeala/GPT4-x-AlpacaDente2-30b</a>	main	57.2	56.1	79.8	44	49.1
<a href="#">TheBloke/dromedary-65b-lora-HF</a>	main	57	57.8	80.8	50.8	38.8
<a href="#">TheBloke/Wizard-Vicuna-13B-Uncensored-HF</a>	main	57	53.6	79.6	42.7	52
<a href="#">elinass/llama-30b-hf-transformers-4.29</a>	main	56.9	57.1	82.6	45.7	42.3
<a href="#">ausboss/llama30b-SuperHOT</a>	main	56.9	57.1	82.6	45.7	42.3

## Modelos adaptados

- Actualmente hay más de 140 modelos comparados.
- Es menos costoso que entrenar modelos base

[https://huggingface.co/spaces/HuggingFaceH4/open\\_llm\\_leaderboard](https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard)

# Modelos de lenguaje

- Ecosistema LLMs: legislación
  - 22 requisites for foundation model providers

Category	Keyword	Requirement (summarized)	Section
Data	Data sources	Describe data sources used to train the foundation model.	Amendment 771, Annex VIII, Section C, page 348
	Data governance	Use data that is subject to data governance measures (suitability, bias, and appropriate mitigation) to train the foundation model.	Amendment 399, Article 28b, page 200
	Copyrighted data	Summarize copyrighted data used to train the foundation model.	Amendment 399, Article 28b, page 200
Compute	Compute	Disclose compute (model size, computer power, training time) used to train the foundation model.	Amendment 771, Annex VIII, Section C, page 348
	Energy	Measure energy consumption and take steps to reduce energy use in training the foundation model.	Amendment 399, Article 28b, page 200
Model	Capabilities/limitations	Describe capabilities and limitations of the foundation model.	Amendment 771, Annex VIII, Section C, page 348
	Risks/mitigations	Describe foreseeable risks, associated mitigations, and justify any non-mitigated risks of the foundation model.	Amendment 771, Annex VIII, Section C, page 348 and Amendment 399, Article 28b, page 200
	Evaluations	Benchmark the foundation model on public/industry standard benchmarks.	Amendment 771, Annex VIII, Section C, page 348 and Amendment 399, Article 28b, page 200
	Testing	Report the results of internal and external testing of the foundation model.	Amendment 771, Annex VIII, Section C, page 348 and Amendment 399, Article 28b, page 200
Deployment	Machine-generated content	Disclose content from a generative foundation model is machine-generated and not human-generated.	Amendment 101, Recital 60g, page 76
	Member states	Disclose EU member states where the foundation model is on the market.	Amendment 771, Annex VIII, Section C, page 348
	Downstream documentation	Provide sufficient technical compliance for downstream compliance with the EU AI Act.	Amendment 101, Recital 60g, page 76 and Amendment 399, Article 28b, page 200

# Modelos de lenguaje

- Ecosistema LLMs: legislación

## Grading Foundation Model Providers' Compliance with the Draft EU AI Act

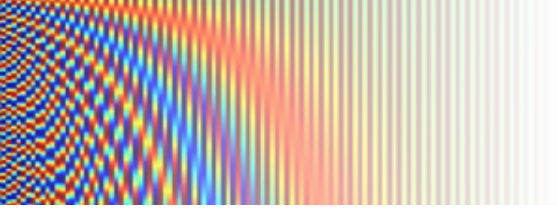
Source: Stanford Research on Foundation Models (CRFM), Institute for Human-Centered Artificial Intelligence (HAI)

	OpenAI	cohere	stability.ai	ANTHROPIC	Google	BigScience	Meta	AI21labs	ALEPH ALPHA	ELEUTHERAI	Totals
Draft AI Act Requirements	GPT-4	Cohere Command	Stable Diffusion v2	Claude	PaLM 2	BLOOM	LLaMA	Jurassic-2	Luminous	GPT-NeoX	
Data sources	●●●●	●●●●	●●●●	○○○○	●●●●	●●●●	●●●●	○○○○	○○○○	●●●●	22
Data governance	●●●●	●●●●	●●○○	○○○○	●●●●	●●●●	●●○○	○○○○	○○○○	●●●●	19
Copyrighted data	○○○○	○○○○	○○○○	○○○○	○○○○	●●●●	○○○○	○○○○	○○○○	●●●●	7
Compute	○○○○	○○○○	●●●●	○○○○	○○○○	●●●●	●●●●	○○○○	●○○○	●●●●	17
Energy	○○○○	●○○○	●●●●	○○○○	○○○○	●●●●	●●●●	○○○○	○○○○	●●●●	16
Capabilities & limitations	●●●●	●●●●	●●●●	●○○○	●●●●	●●○○	●●○○	●●○○	●○○○	●●●●	27
Risks & mitigations	●●●●	●●○○	●○○○	●○○○	●●●●	●●○○	●○○○	●●○○	○○○○	●○○○	16
Evaluations	●●●●	●●○○	○○○○	○○○○	●○○○	●●●●	●○○○	○○○○	●○○○	●○○○	15
Testing	●●●●	●●○○	○○○○	○○○○	●○○○	●●○○	○○○○	●○○○	○○○○	○○○○	10
Machine-generated content	●●●●	●●●●	○○○○	●●●●	●●●●	●●●●	○○○○	●●●●	●○○○	●●○○	21
Member states	●●●●	○○○○	○○○○	●●○○	●●●●	○○○○	○○○○	○○○○	●○○○	○○○○	9
Downstream documentation	●●●●	●●●●	●●●●	○○○○	●●●●	●●●●	●●○○	○○○○	○○○○	●●●●	24
Totals	25 / 48	23 / 48	22 / 48	7 / 48	27 / 48	36 / 48	21 / 48	8 / 48	5 / 48	29 / 48	

@stanfordCRFM

13 junio European Parliament approved the draft AI Act.

<https://crfm.stanford.edu/2023/06/15/eu-ai-act.html>



# Casos de uso

Casos de uso

# Casos de uso

- Transformers, ejemplos
  - Empezamos por ejemplos de mal uso
    - Muchas veces generan **información incompleta, inexacta o falsa** aunque la apariencia de seguridad del texto nos hace creer lo contrario (**alucinaciones**)
    - No es buena práctica basar toda la conversación en torno a un acceso a **hechos o datos concretos** como si fuera una enciclopedia
      - Un abogado presentó como jurisprudencia **un caso que se inventó chatgpt**

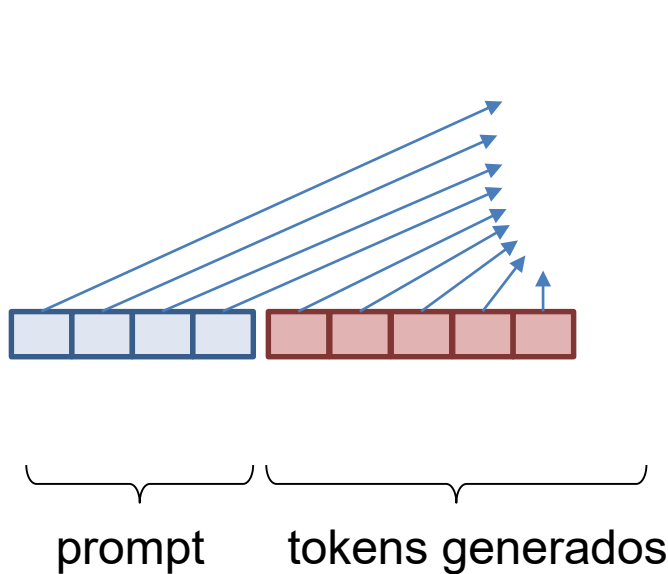
**B B C**

**ChatGPT: US lawyer admits using AI  
for case research**

27/05/2023

# Casos de uso

- Generación



modelo usa la siguiente **información**:

- almacenada en sus **pesos: poco fiable**
- **prompt** tenemos control
- tokens generados previamente
- en chats **turnos anteriores**

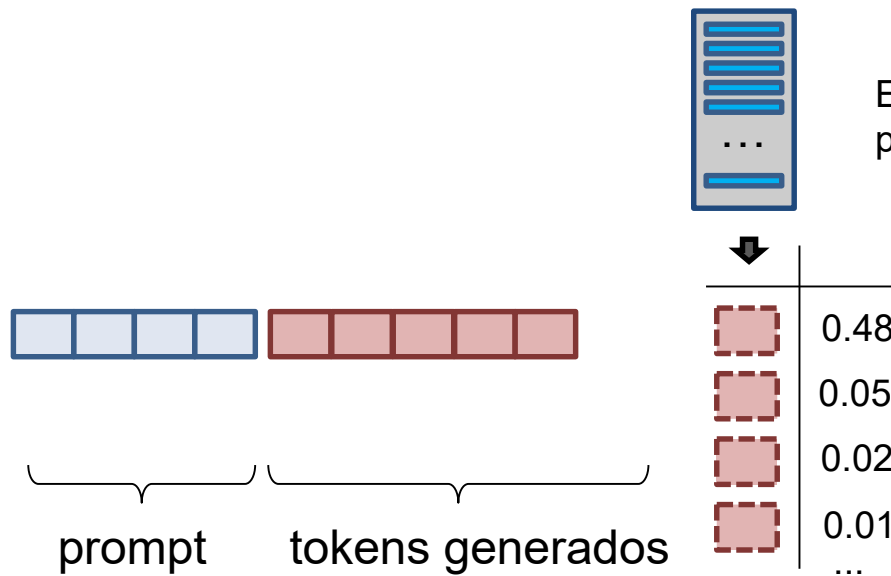
} self-attention

El **diseño del prompt** y la inclusión de información válida en él es la forma más rápida de mejorar la calidad

El modelo tiene **acceso inmediato a toda la entrada**, por eso hay una demanda por ampliar la capacidad de los prompt: 16k, 32k en gpt3, y gpt4

# Casos de uso

- Generación

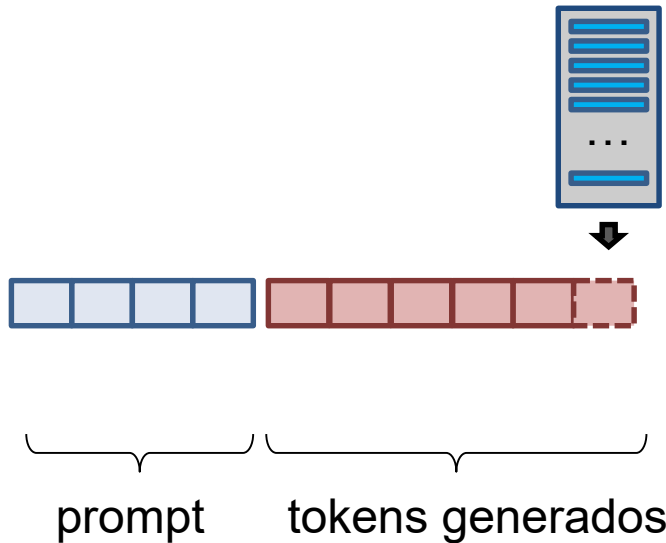


El **nuevo token** se obtiene por muestreo de las probabilidades de salida



# Casos de uso

- Generación



Si se produce un error (al muestrear) sigue con él. Hay soluciones

Sin acceso a la programación se pueden utilizar estrategias en el prompt como

- Dar varios ejemplos **few shot** así la atención usará los ejemplos para buscar parecidos con lo que se pregunta
- Pedir explícitamente en el prompt que se desarrolle **paso a paso**, incluso impersonando una persona muy cualificada
- Utilizar la atención de la conversación para que repase su salida anterior, con los requisitos **preguntando si su respuesta anterior ha cumplido**

Ejemplos:

- Let's work this out in a step by step way to be sure we have the right answer (82%)
- Let's think step by step. (78%)
- zero shot (17%)

Arithmetic Reasoning

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: **Let's think step by step.**

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

*Kojima, T., et al (2022). Large language models are zero-shot reasoners. arXiv preprint arXiv:2205.11916.*

*Zhou, Y., et al (2022). Large language models are human-level prompt engineers. arXiv preprint arXiv:2211.01910.*

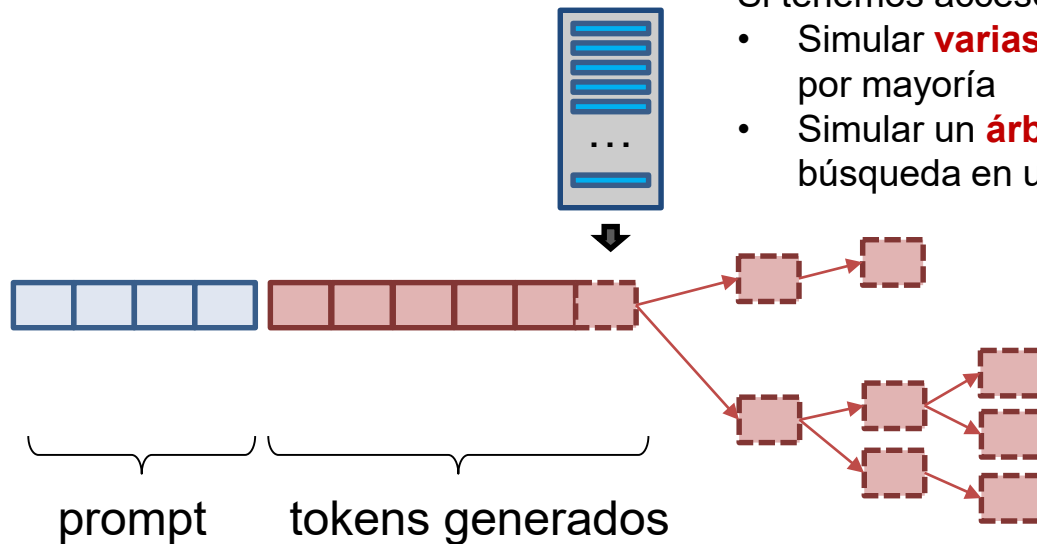
# Casos de uso

- Generación

Si se produce un error (al muestrear) sigue con él. Hay soluciones

Si tenemos acceso al desarrollo

- Simular **varias cadenas de pensamientos** y elegir por mayoría
- Simular un **árbol de pensamientos**, como una búsqueda en un juego.



*Yao, S et al (2023). Tree of thoughts: Deliberate problem solving with large language models. arXiv preprint arXiv:2305.10601. (17 mayo 2023, 7 citas)*



# Casos de uso

- Transformers, ejemplos: uso adecuado
  - Siempre que sea una tarea que sabemos hacer y podemos **verificar** lo vamos a ver como una **ayuda** a la productividad, **inspiración**
    - Resumir texto, responder a preguntas



# Casos de uso

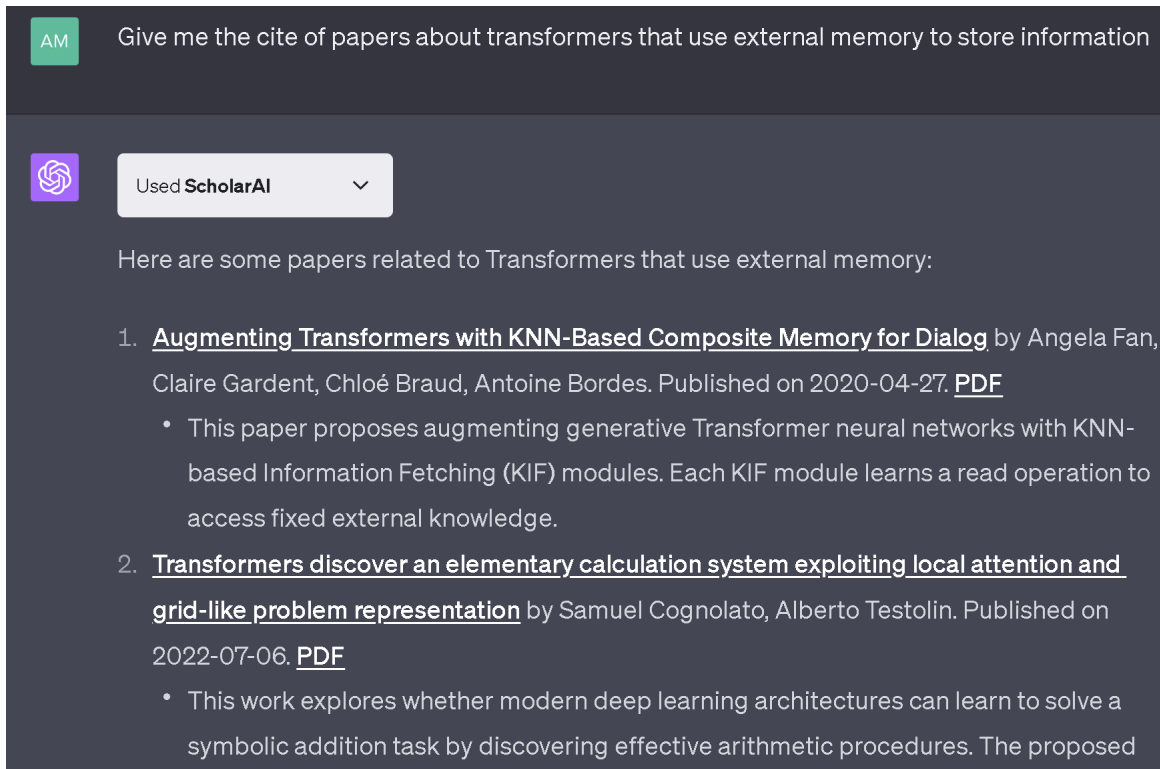
- Transformers, ejemplos: uso adecuado
  - Siempre que sea una tarea que sabemos hacer y podemos **verificar** lo vamos a ver como una **ayuda** a la productividad, **inspiración**
    - Resumir texto, responder a preguntas, asistente programación
  - Este tipo de uso además se puede complementar con el hecho de que sea una conversación
    - Tenemos que asumir que no necesitamos respuestas correctas a la primera, **podemos refinar mediante la conversación**

# Casos de uso

- Transformers, ejemplos: uso adecuado
  - Siempre que sea una tarea que sabemos hacer y podemos **verificar** lo vamos a ver como una **ayuda** a la productividad, **inspiración**
    - Resumir texto, responder a preguntas
  - Este tipo de uso además se puede complementar con el hecho de que sea una conversación
    - Tenemos que asumir que no necesitamos respuestas correctas a la primera, **podemos refinar mediante la conversación**
  - Podemos **alimentar el prompt con datos correctos** para que trabaje sobre ellos
    - Esto es lo que hacen algunos del ecosistema de plug-ins que han surgido alrededor de estos asistentes
      - Aportan contacto con el mundo real (no una versión congelada en unos parámetros)
      - Conexión a bases de datos, bibliografía, imágenes, mapas

# Casos de uso

- Transformers, ejemplos, contenido fiable (plugin Scholar AI)



AM Give me the cite of papers about transformers that use external memory to store information

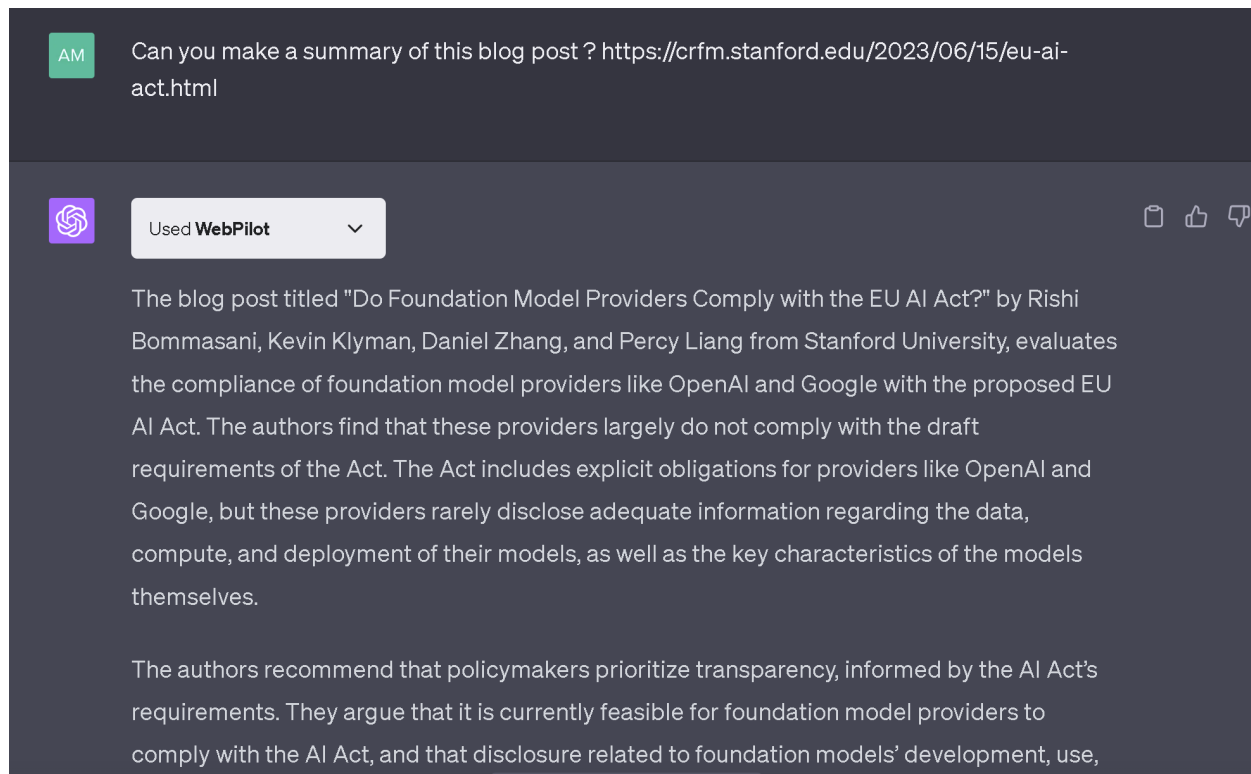
Used **ScholarAI**

Here are some papers related to Transformers that use external memory:


1. [Augmenting Transformers with KNN-Based Composite Memory for Dialog](#) by Angela Fan, Claire Gardent, Chloé Braud, Antoine Bordes. Published on 2020-04-27. [PDF](#)
  - This paper proposes augmenting generative Transformer neural networks with KNN-based Information Fetching (KIF) modules. Each KIF module learns a read operation to access fixed external knowledge.
2. [Transformers discover an elementary calculation system exploiting local attention and grid-like problem representation](#) by Samuel Cognolato, Alberto Testolin. Published on 2022-07-06. [PDF](#)
  - This work explores whether modern deep learning architectures can learn to solve a symbolic addition task by discovering effective arithmetic procedures. The proposed

# Casos de uso

- Transformers, ejemplos, contenido externo (plugin web pilot)



AM Can you make a summary of this blog post ? <https://crfm.stanford.edu/2023/06/15/eu-ai-act.html>

Used **WebPilot** 

The blog post titled "Do Foundation Model Providers Comply with the EU AI Act?" by Rishi Bommasani, Kevin Klyman, Daniel Zhang, and Percy Liang from Stanford University, evaluates the compliance of foundation model providers like OpenAI and Google with the proposed EU AI Act. The authors find that these providers largely do not comply with the draft requirements of the Act. The Act includes explicit obligations for providers like OpenAI and Google, but these providers rarely disclose adequate information regarding the data, compute, and deployment of their models, as well as the key characteristics of the models themselves.

The authors recommend that policymakers prioritize transparency, informed by the AI Act's requirements. They argue that it is currently feasible for foundation model providers to comply with the AI Act, and that disclosure related to foundation models' development, use,



# Casos de uso

- Transformers, asistentes a la programación
  - Github en colaboración con openAI 1 jul 2021

```
1 import datetime
2
3 def parse_expenses(expenses_string):
4     """Parse the list of expenses and return the list of triples (date, value, currency).
5     Ignore lines starting with #.
6     Parse the date using datetime.
7     Example expenses_string:
8         2016-01-02 -34.01 USD
9         2016-01-03 2.59 DKK
10        2016-01-03 -2.72 EUR
11    """
12    expenses = []
13    for line in expenses_string.splitlines():
14        if line.startswith("#"):
15            continue
16        date, value, currency = line.split(" ")
17        expenses.append((datetime.datetime.strptime(date, "%Y-%m-%d"),
18                        float(value),
19                        currency))
20    return expenses
```

```
# this function trains a decision tree model for the given training data and labels
def train_model(train_data, train_labels):
    # TODO: train the decision tree model
    # - create a DecisionTreeClassifier object, fit it to the training data and labels
    # - return the object
    clf = tree.DecisionTreeClassifier()
    clf = clf.fit(train_data, train_labels)
    return clf

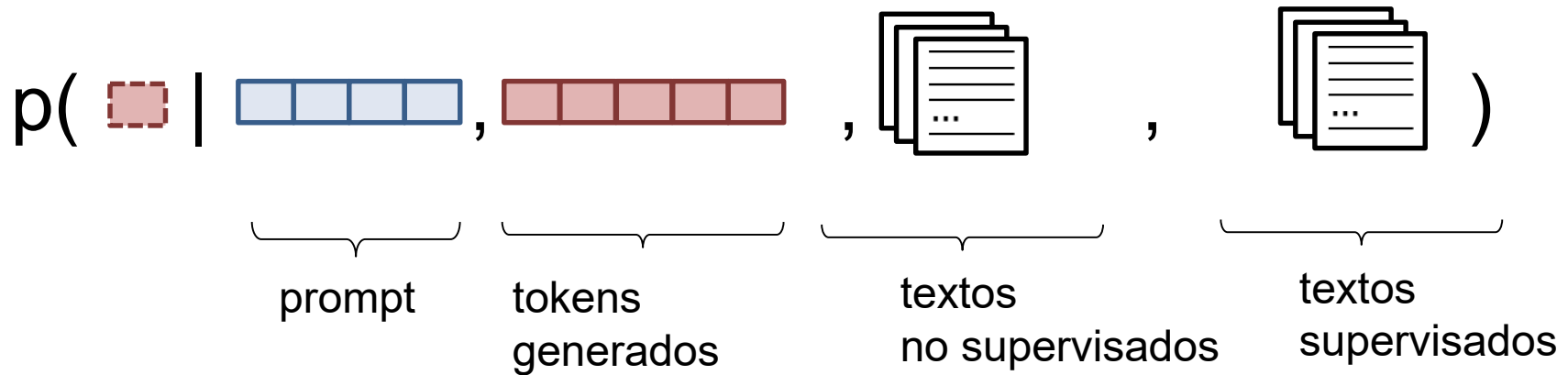
# this function compares three vectors with cosine similarity and returns the most similar pair
def most_similar(vector_1, vectors):
    # TODO: find the pair of vectors that are most similar
    # - create a list of the cosine similarities between vector_1 and vectors
    cos_sim = []
    for vector in vectors:
        cos_sim.append(cosine_similarity(vector_1, vector))
    # - sort the list in descending order
    cos_sim.sort(reverse=True)
    # - return the pair of vectors that are most similar
    return vectors[cos_sim.index(cos_sim[0])]
```

<https://github.com/features/copilot>

# Casos de uso

- Sesgo /Bias

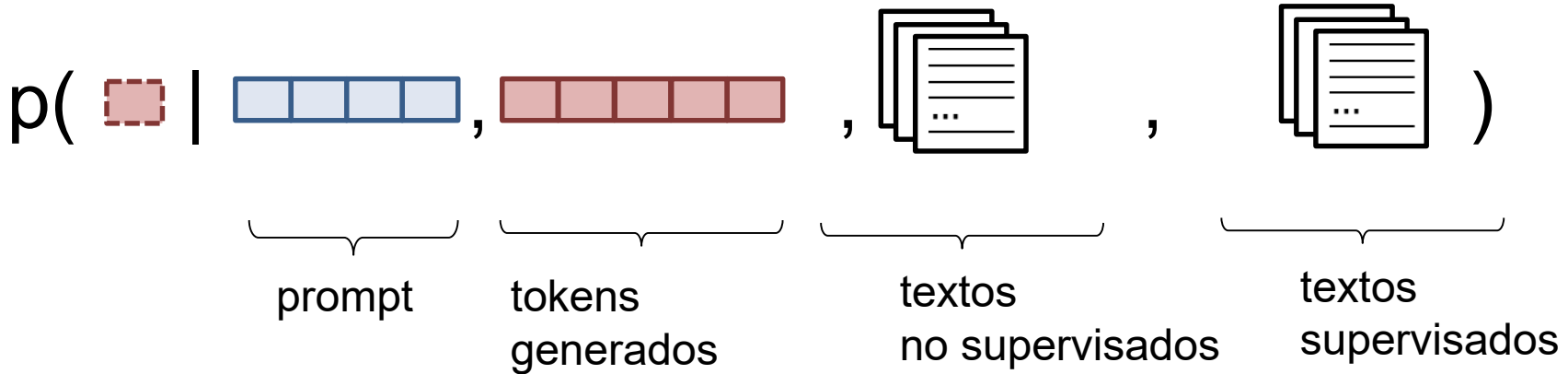
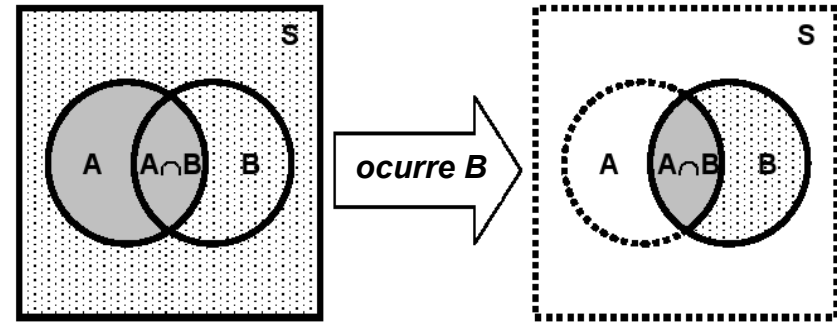
La función de densidad de probabilidad para cada token:



# Casos de uso

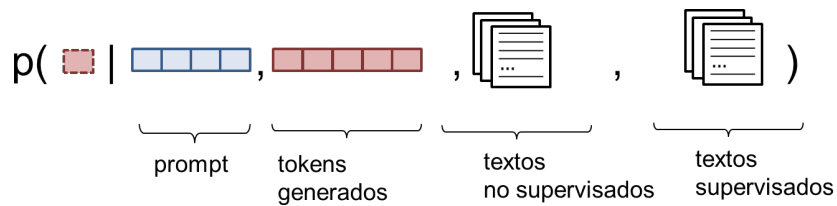
- Sesgo /Bias

Con el contexto **condicionamos** es espacio sobre el que muestrear



# Casos de uso

- Sesgo /Bias



'The patient was ready for the surgery. The doctor took the gloves.

llama7b

$p(\text{'He'})$  0.20

...

$p(\text{'She'})$  0.01

Let us write avoiding job related gender bias, from now on I will write he and she with equal chances of the charactes of being male of female independently of their job. For example the text:

The patient was ready for the surgery. The doctor took the gloves. `

llama7b

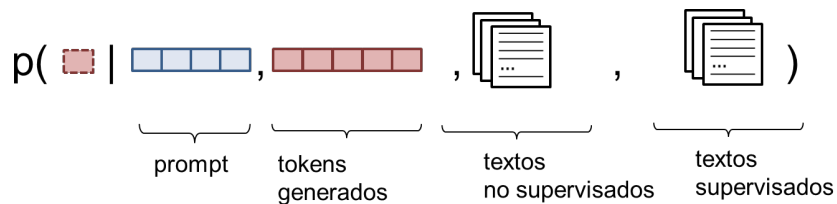
$p(\text{'He'})$  0.11

...

$p(\text{'She'})$  0.03

# Casos de uso

- Sesgo /Bias



'The patient was ready for the surgery. The doctor took the gloves. '

gpt3-turbo  
male: 65%  
female: 10%

Let us write avoiding job related gender bias, from now on I will write he and she with equal chances of the charactes of being male of female independently of their job. For example the text:

The patient was ready for the surgery. The doctor took the gloves. '

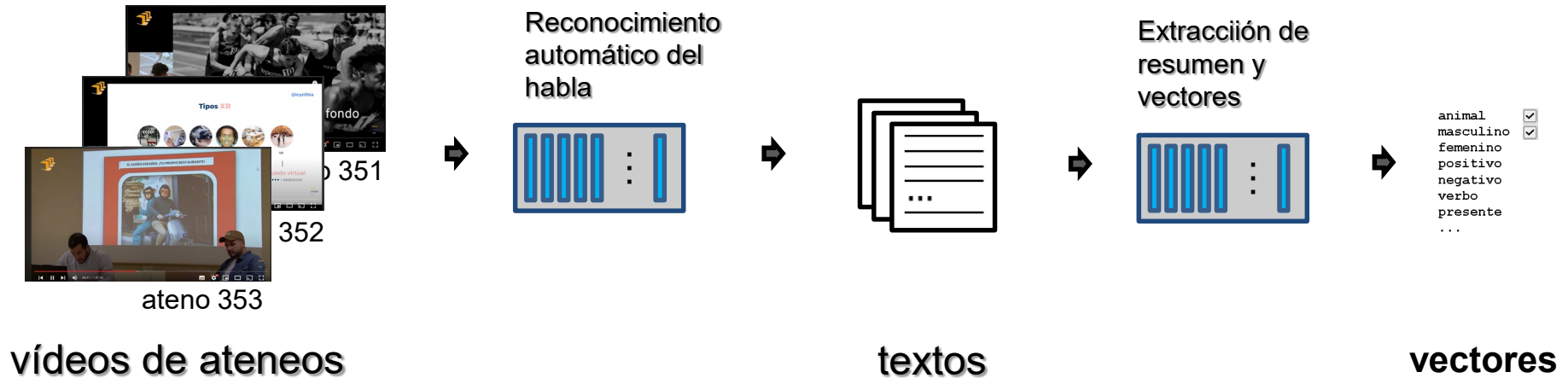
gpt3-turbo  
male: 45%  
female: 15%  
neutral:40% (He/She or they)

# Casos de uso

- Transformers: ejemplo asistente sobre Ateneos

1. Transcribimos los ateneos mediante un sistema estado del arte (basado en transformers)
2. Obtenemos representaciones vectoriales de los textos y su resumen (transformer)
3. Al crear resúmenes se corrigien muchos errores del reconocedor por el contexto

...



# Casos de uso

- Transformers: ejemplo asistente sobre Ateneos

Qué oportunidades se han presentado relacionadas con el coche de hidrógeno.

-----

OPORTUNIDADES DE LA TECNOLOGÍA DEL HIDRÓGENO COMO VECTOR ENERGÉTICO. Fernando Palacín Arizón. 25/10/2017  
Las oportunidades que se presentan relacionadas con el coche de hidrógeno incluyen una mayor autonomía en comparación con los coches eléctricos, un tiempo de carga más rápido de 2-3 minutos al recargar el depósito de hidrógeno en lugar de la batería, y una reducción en la necesidad de mantenimiento. Además, se menciona que los fabricantes de automóviles están desarrollando coches de hidrógeno y se espera que mejoren en términos de prestaciones y costos en los próximos años.

BMW I, LA ECONOMÍA CIRCULAR APLICADA A LA AUTOMOCIÓN. Leonardo Santamaría Hernández. 16/05/2018  
El texto menciona que la compañía tiene un acuerdo de intercambio con Toyota en el que les proporcionaron motores y ellos proporcionaron tecnología para desarrollar nuevos modelos con hidrógeno. Aunque no se proporcionan detalles específicos sobre las oportunidades relacionadas con el coche de hidrógeno, se menciona que en el futuro es posible que cambien a vehículos de hidrógeno.

LA ENERGÍA: UNA OPORTUNIDAD PARA EL DESARROLLO ECONÓMICO. Rafael Moliner. 29/04/2015  
En la charla se menciona que el coche de hidrógeno es una alternativa a la crisis del petróleo y al cambio climático. Se habla sobre tecnologías existentes para implementar la economía del hidrógeno en el transporte. Se menciona que se puede producir hidrógeno a partir del carbono sin emitir CO2 y a precios más competitivos. También se menciona la posibilidad de almacenar electricidad en pilas en lugar de producir hidrógeno por electrólisis. Se afirma que el futuro de los coches es híbrido, con batería para una autonomía eléctrica de 200-300 km y una batería de combustible para mayor autonomía. Además, se habla del desarrollo actual de laboratorio que sugiere que se puede duplicar la capacidad de los supercondensadores.



# Casos de uso

- Transformers: ejemplo asistente sobre Ateneos

¿ Qué oportunidades se presentan para Teruel en el campo de la aeronáutica ?

---

En el campo de la aeronáutica, Teruel tiene oportunidades en el negocio de reciclaje de aviones, ya que se espera que se reciclen más de 15,000 aviones en la zona, lo que representa el 10% de todas las aeronaves del mundo. Además, la creación de un 30% más de rutas aéreas en los próximos 20 años representa una ventaja para este negocio en Teruel. También se destaca que el aeropuerto de Teruel es considerado maravilloso y tiene tasas baratas, lo que atrae a la aviación ejecutiva o general desde varios lugares de España y de otros países europeos. Por lo tanto, hay oportunidades para el crecimiento y desarrollo en el campo de aeronáutica en Teruel.

# Casos de uso

- Transformers: ejemplo asistente sobre Ateneos

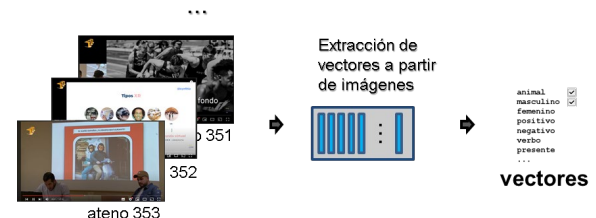
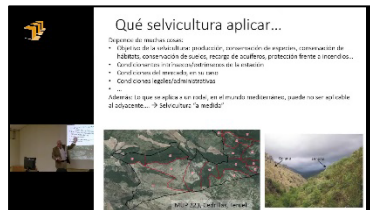
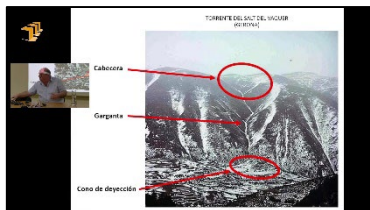
¿Cuáles son los retos más importantes a los que se enfrenta el Aragonés como lengua?

-----

Los retos más importantes a los que se enfrenta el aragonés como lengua son la estigmatización y la falta de transmisión intergeneracional. El hablante menciona que los hablantes tienen derecho a abandonar su lengua, pero esto a menudo ocurre debido a la estigmatización que rodea al aragonés. Además, la falta de transmisión de padres a hijos es crucial para la supervivencia de la lengua. Por lo tanto, es importante que la lengua se viva y se transmita a las nuevas generaciones, y que se aprecie como parte del patrimonio lingüístico de Aragón. También se menciona la importancia de construir una estrategia de política lingüística a largo plazo para el aragonés, basada en el consenso y la participación de la comunidad. Esto implica promover el aragonés de una manera práctica y cotidiana, así como utilizarlo como una estrategia lingüística en lugar de centrarse solo en su estudio filológico.

# Casos de uso

## • Transformers: ejemplo asistente sobre Ateneos



'slide with a landscape picture of mountain village'

- Ateneo 340, título/autor/año: DESAFÍOS DEL CAMBIO CLIMÁTICO: IMPACTOS, MITIGACIÓN Y ADAPTACIÓN. José M<sup>a</sup> Cudrat Prats. 26/10/22 .

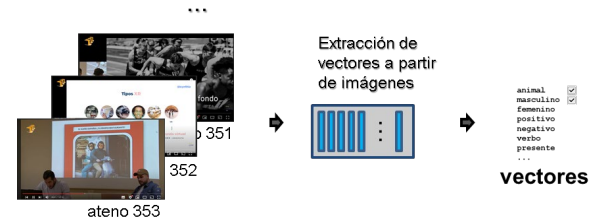
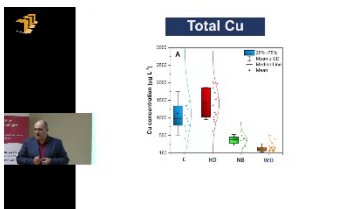
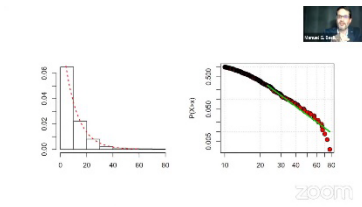
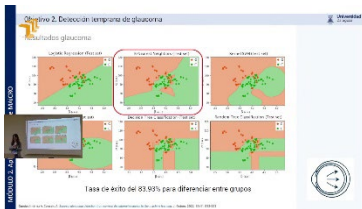
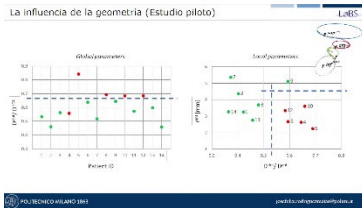
- Ateneo 318, título/autor/año: LOS PINOS NO SON HIJOS DEL FRANQUISMO: SU "MALA PRENSA" ES UN MITO. Luis Alfonso Gil Sánchez. 17/02/21

- Ateneo 341, título/autor/año: EL AGUA NO SALE DE LOS RÍOS:HAGAMOS RESTAURACIÓN HIDROLÓGICO-FORESTAL DE CU. Ignacio Pérez-Soba Diez del Corral. 9/11/22

- Ateneo 337, título/autor/año: AL MARGEN DE PREJUICIOS URBANOS: CORTAR NO ES PERJUDICAR. Miguel Cabrera. 27/04/22 . data/es/ateneo/YC4m\_ss07h8/YC4m\_ss07h8.img/r200215.jpg

# Casos de uso

- Transformers: ejemplo asistente sobre Ateneos



'slide with a results graph plot'

- Ateneo 347, título/autor/año: EXPERIENCIAS DE MEDICINA in-silico. José Félix Rodríguez Matas. 1/03/23 .

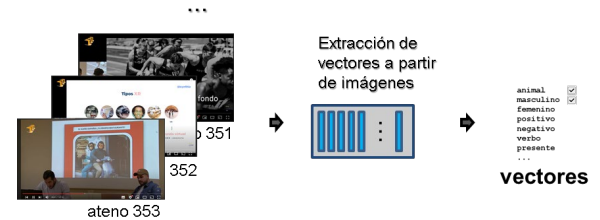
- Ateneo 346, título/autor/año: MAiCRO: CÓMO DETECTAR DAÑO ACOLAR CON UNA FOTOGRAFÍA. Alejandro Consejo. 15/02/23 .

- Ateneo 316, título/autor/año: INVESTIGANDO SOBRE LA INVESTIGACIÓN: MÁS ALLÁ DEL TALENTO. Manuel G. Bedia. 20/01/2021 .

- Ateneo 350, título/autor/año: MICROMUESTREO, ESPECTROMETRÍA ATÓMICA E INTELIGENCIA ARTIFICIAL, NUEVAS POSIBILIDADES PARA EL DIAGNÓSTICO MÉDICO. Martín Resano. 19/04/23 .

# Casos de uso

- Transformers: ejemplo asistente sobre Ateneos



'image with a person standing next to a slide with hands raised'

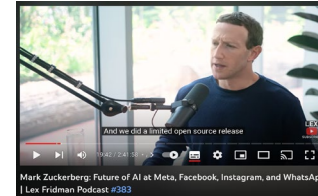
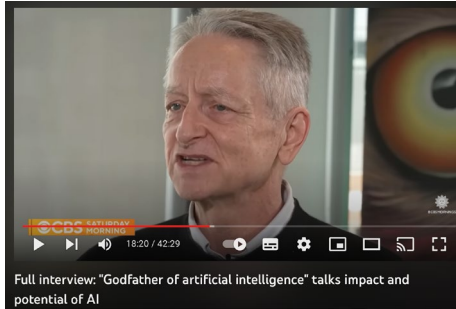
- Ateneo 275, título/autor/año: LA AERONAÚTICA DESDE UNA NUEVA PERSPECTIVA DE GÉNERO.  
Alejandro Ibrahim Perera. 01/02/2017. data/es/ateneo/90rMFE6Fi6M/90rMFE6Fi6M.img/r200154.jpg

- Ateneo 260, título/autor/año: ATAPUERCA Y EVOLUCIÓN HUMANA: ¿CÓMO SABEMOS QUIÉN ES QUIÉN Y CUÁL ES SU ANTIGÜEDAD?.  
Gloria Cuenca Bescós. 18/11/2015.

- Ateneo 306, título/autor/año: EMPRENDIMIENTO EN LA ERA DE LA TRANSFORMACIÓN DIGITAL.  
Ana Monreal Vidal. 4/12/2019

- Ateneo 329, título/autor/año: SERVICIOS E INFRAESTRUCTURA DE AWS Y LA PRÓXIMA REGIÓN EN ARAGÓN.  
Javier Ramírez. 15/12/21 .

# Conclusiones



- El miedo que plantea surge por la **velocidad** a la que está sucediendo todo
- Prácticamente en 4 años hemos replanteado el concepto de **complejidad**, pasando de considerar complicado extraer entidades de un texto o deshacer ambigüedades a estar haciendo resúmenes y respondiendo preguntas hoy en día
- Estos modelos nos asustan por cómo **representan la información** (de una manera más cercana a la nuestra)
- El debate está abierto sobre la necesidad de **regulación** ¿y si no lo hacen todos los países?
- La **UE** se ha lanzado a regular y amenazar con multas
- En Japón han declarado que todo el **copyright** no aplica a modelos de aprendizaje
- Algunas empresas y entidades apuestan por el **software libre** y reconocer haber utilizado desarrollos recientes como llama.cpp en su flujo de trabajo





Gracias por la atención ¿ alguna pregunta ?

