

Summit País Digital
Hackatón by EY
colabora Microsoft

Large Language Models & Prompting

Septiembre 2023
Sesión 2



The better the question. The better the answer.
The better the world works.

HA
CKA
TÓN

XI
20
23

summit
PAÍS
DIGITAL



By:



Colabora:



Calendario de Capacitaciones

El lenguaje de programación base para todas las capacitaciones y esta hackatón es “Python”. El calendario de las capacitaciones es el siguiente, este será comunicado a los participantes de la hackatón y puede estar sujeto a cambios por situaciones de fuerza mayor:

- ~~Sesión 1: martes 12 de septiembre de 19:00 a 20:15~~ horas de Chile - “Design Thinking”
- Sesión 2: miércoles 13 de septiembre de 19:00 a 20:15 horas de Chile - “LLM & Prompting”
- Sesión 3: jueves 14 de septiembre de 19:00 a 20:15 horas de Chile - “APIS de OpenAI”
- Sesión 4: martes 26 de septiembre de 19:00 a 20:15 horas de Chile - “Langchain”
- Sesión 5: miércoles 27 de septiembre de 19:00 a 20:15 horas de Chile - “Casos de Uso / Wireframing”
- **IMPORTANTE** Sesión 6 Obligatoria (asistencia por lo menos de dos participantes por equipo): jueves 28 de septiembre 2023 de 19:00 a 20:15 horas de Chile - “Pitch Class y Aclaratorias generales”

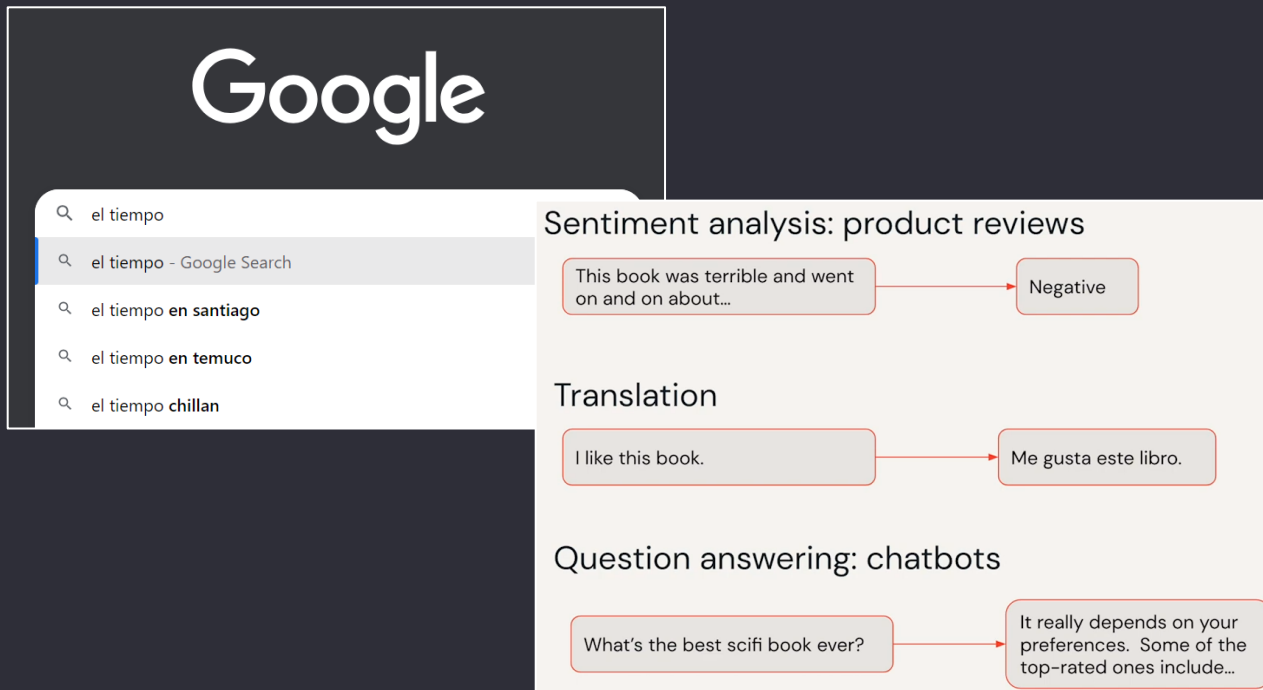
Todas las sesiones de capacitación quedarán grabadas y estarán disponibles en línea para los participantes del evento.

Agenda

- 1 Introducción a LLMs
- 2 Prompt y Prompt Engineering
- 3 Técnicas de Prompting
- 4 Recomendaciones finales

Antes que LLM, ¿qué es NLP?

- ▶ Natural Language Processing (NLP) es una rama de la inteligencia artificial que le brinda a los computadores la capacidad de interpretar, manipular y comprender el lenguaje humano.
- ▶ De esta manera, los computadores pueden volverse eficientes en tareas como la traducción, análisis de sentimiento, reconocimiento de voz y la generación de texto, video e imágenes.
- ▶ ¡Usamos NLP todos los días!



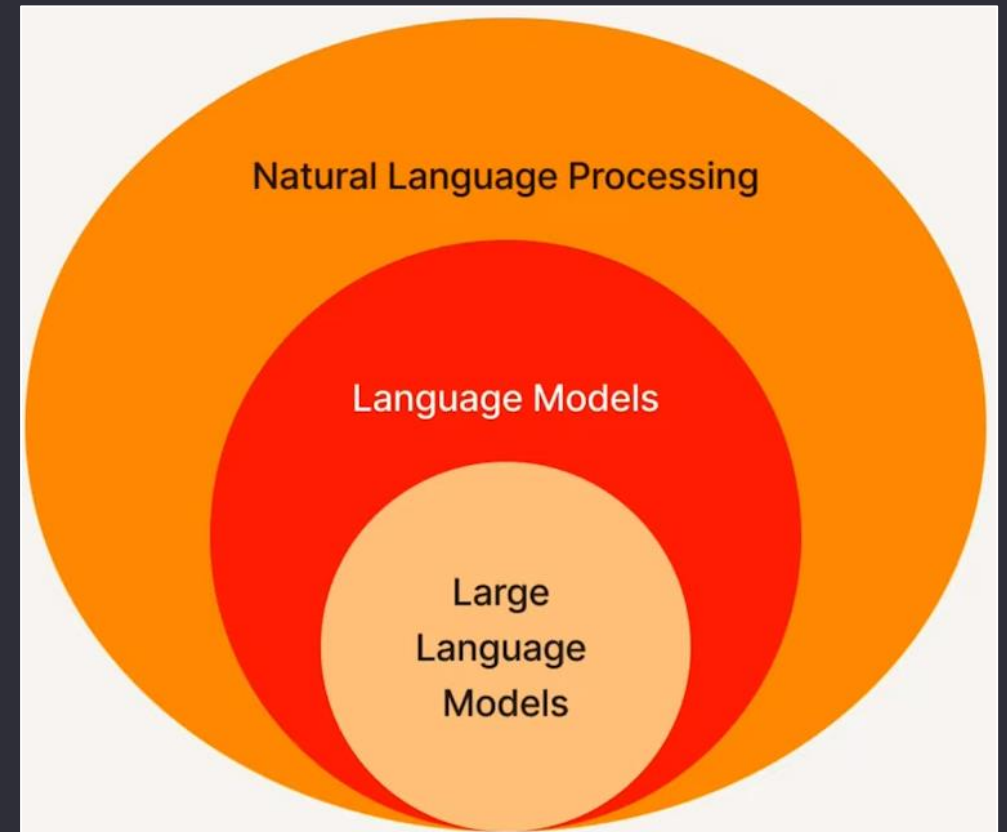
The image is a composite showing a Google search interface on the left and three NLP application examples on the right.

Google Search Interface:

- Search bar: el tiempo
- Suggestion 1: el tiempo - Google Search
- Suggestion 2: el tiempo en santiago
- Suggestion 3: el tiempo en temuco
- Suggestion 4: el tiempo chillan

NLP Applications:

- Sentiment analysis: product reviews**
Input: "This book was terrible and went on and on about..." → Output: Negative
- Translation**
Input: "I like this book." → Output: Me gusta este libro.
- Question answering: chatbots**
Input: "What's the best scifi book ever?" → Output: It really depends on your preferences. Some of the top-rated ones include...



¿Cómo leen texto los computadores?

Conceptos básicos de NLP

- ▶ **Tokenización:** Proceso de dividir un texto en unidades individuales (tokens), que luego se utilizan para análisis o procesamiento adicional.
- ▶ **Token:** Unidad de lenguaje que ha sido extraída durante un determinado proceso de tokenización
- ▶ **Secuencia:** Serie ordenada de tokens
- ▶ **Vocabulario:** Conjunto de palabras únicas (o tokens) presentes en un conjunto de datos específico

"The moon, Earth's only natural satellite, has been a subject of fascination and wonder for thousands of years"			
Token Basic building block <ul style="list-style-type: none">• The• Moon• ,• Earth's• Only•• years	Sequence Sequential list of tokens <ul style="list-style-type: none">• The moon,• Earth's only natural satellite• Has been a subject of•• Thousands of years	Vocabulary Complete list of tokens { 1: "The", 569: "moon", 122: ",", 430: "Earth", 50: "'s", ...}	
Tokenization method	Tokens	Token count	Vocab size
Sentence	'The moon, Earth's only natural satellite, has been a subject of fascination and wonder for thousands of years.'	1	# sentences in doc
Word	'The', 'moon,', 'Earth's', 'only', 'natural', 'satellite,', 'has', 'been', 'a', 'subject', 'of', 'fascination', 'and', 'wonder', 'for', 'thousands', 'of', 'years.'	18	171K (English ¹)
Sub-word	'The', 'moon', ',', 'Earth', "'", 's', 'on', 'ly', 'n', 'atur', 'al', 's', 'ate', 'll', 'it', 'e', ',', 'has', 'been', 'a', 'subject', 'of', 'fascinat', 'ion', 'and', 'w', 'on', 'd', 'er', 'for', 'th', 'ous', 'and', 's', 'of', 'y', 'ears', ','	37	(varies)
Character	'T', 'h', 'e', ',', 'm', 'o', 'o', 'n', ',', 'E', 'a', 'r', 't', 'h', "'", 's', ',', 'o', 'n', 'l', 'y', ',', 'h', 'a', 's', 'b', 'e', 'e', 'n', 'a', 's', 'u', 'b', 'j', 'e', 'c', 't', 'o', 'f', 'f', 'a', 's', 'c', 'i', 'n', 'a', 't', 'i', 'o', 'n', 'a', 'n', 'd', 'w', 'o', 'n', 'd', 'e', 'r', 'f', 'o', 'r', 't', 'h', 'o', 'u', 's', 'a', 'n', 'd', 's', 'o', 'f', 'y', 'e', 'a', 'r', 's', ','	110	52 + punctuation (English)

¿Cómo leen texto los computadores?

► Tokenización por Caracteres

Ventajas

- Vocabulario Reducido
- Robustez a Palabras Nuevas

Desventajas:

- Pérdida de Información Semántica
- Mayor Costo Computacional

► Tokenización por Palabras

Ventajas:

- Simplicidad
- Preservación de la Semántica
- Eficiencia Computacional

Desventajas:

- Vocabulario Extenso
- Manejo de Palabras Nuevas

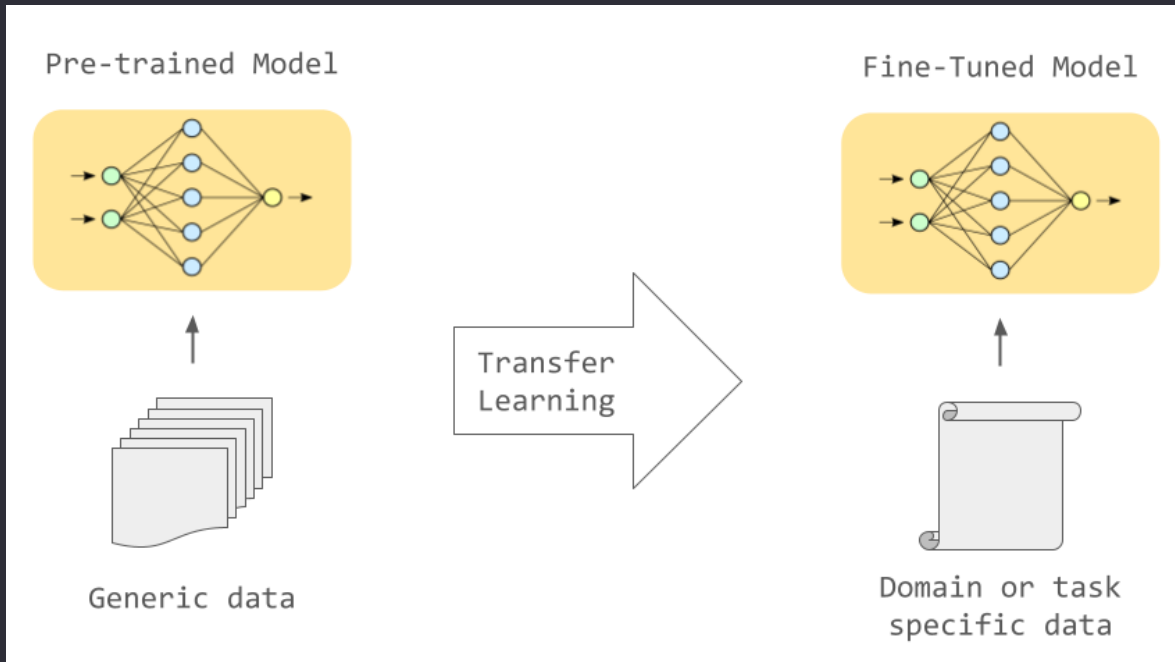
La Tokenización por Sub-Palabras es la metodología más utilizada hoy en día



"The moon, Earth's only natural satellite, has been a subject of fascination and wonder for thousands of years"			
Token	Sequence	Vocabulary	
Basic building block	Sequential list of tokens	Complete list of tokens	
<ul style="list-style-type: none">• The• Moon• ,• Earth's• Only•• years	<ul style="list-style-type: none">• The moon,• Earth's only natural satellite• Has been a subject of•• Thousands of years	{ 1: "The", 569: "moon", 122: ",", 430: "Earth", 50: "'s", ...}	
Tokenization method	Tokens	Token count	Vocab size
Sentence	'The moon, Earth's only natural satellite, has been a subject of fascination and wonder for thousands of years.'	1	# sentences in doc
Word	'The', 'moon,', 'Earth's', 'only', 'natural', 'satellite,', 'has', 'been', 'a', 'subject', 'of', 'fascination', 'and', 'wonder', 'for', 'thousands', 'of', 'years.'	18	171K (English ¹)
Sub-word	'The', 'moon', ',', 'Earth', "'", 's', 'on', 'ly', 'n', 'atur', 'al', 's', 'ate', 'll', 'it', 'e', ',', 'has', 'been', 'a', 'subject', 'of', 'fascinat', 'ion', 'and', 'w', 'on', 'd', 'er', 'for', 'th', 'ous', 'and', 's', 'of', 'y', 'ears', ','	37	(varies)
Character	'T', 'h', 'e', ',', 'm', 'o', 'o', 'n', ',', 'E', 'a', 'r', 't', 'h', "'", 's', "'", 'o', 'n', 'l', 'y', ',', 'n', 'a', 't', 'u', 'r', 'a', 'l', 's', 'a', 't', 'e', 'l', 'l', 'i', 't', 'e', ',', 'h', 'a', 's', ',', 'b', 'e', 'e', 'n', ',', 'a', ',', 's', 'u', 'b', 'j', 'e', 'c', 't', ',', 'o', 'f', ',', 'f', 'a', 's', 'c', 'i', 'n', 'a', 't', 'i', 'o', 'n', ',', 'a', 'n', 'd', ',', 'w', 'o', 'n', 'd', 'e', 'r', ',', 'f', 'o', 'r', ',', 't', 'h', 'o', 'u', 's', 'a', 'n', 'd', ',', 's', ',', 'o', 'f', ',', 'y', 'e', 'a', 'r', 's', ',', '	110	52 + punctuation (English)

¿Y qué es un LM?

Language Models (LMs): Modelos probabilísticos diseñados para identificar y aprender patrones estadísticos en lenguaje natural.



¿Cómo se entrenan estos modelos?

- Mediante aprendizaje auto-supervisado

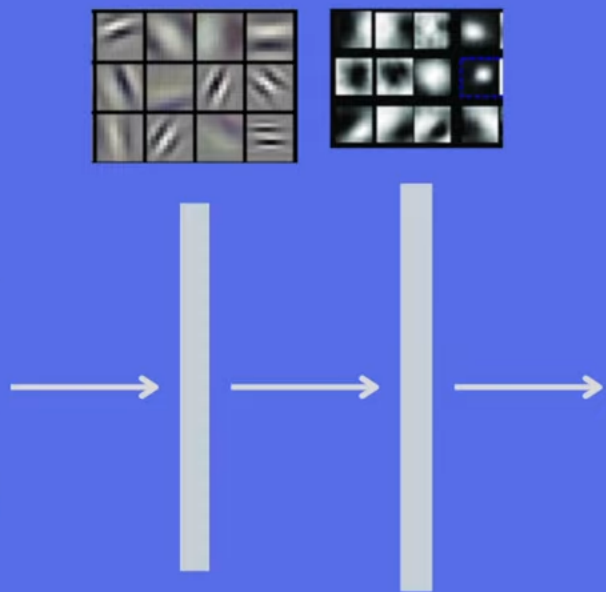
Modelo Pre - Entrenado

modelo crea una representación interna del lenguaje

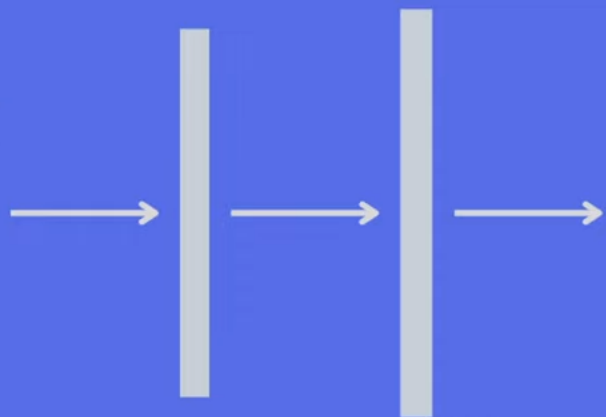
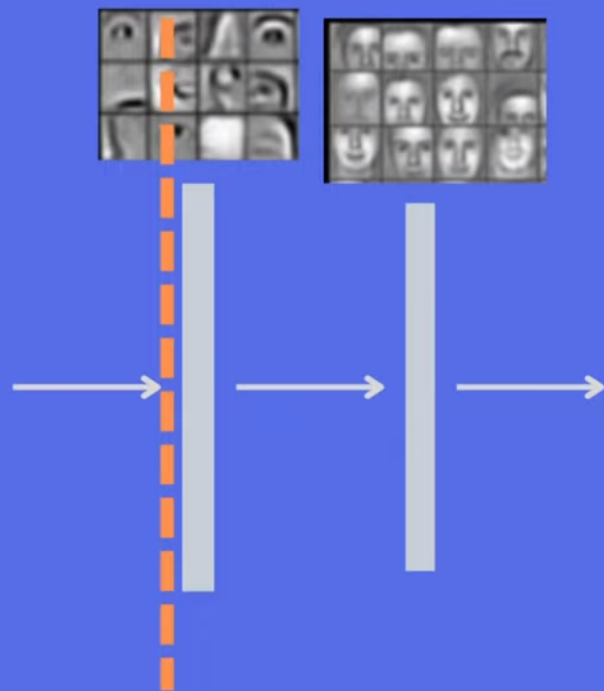
¿Cómo Potenciar un LM Pre Entrenado?

Mediante **Transfer Learning** podemos “tunear” nuestro modelo para realizar tareas específicas

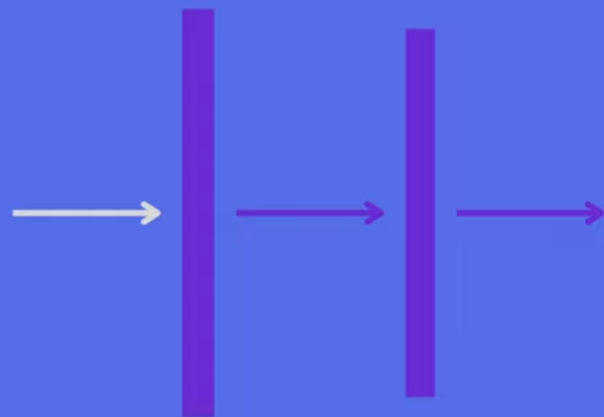
- ¿Qué transferir? (¿qué parte del conocimiento?)
- ¿Cuándo transferir?
- ¿Cómo transferir?



...



...

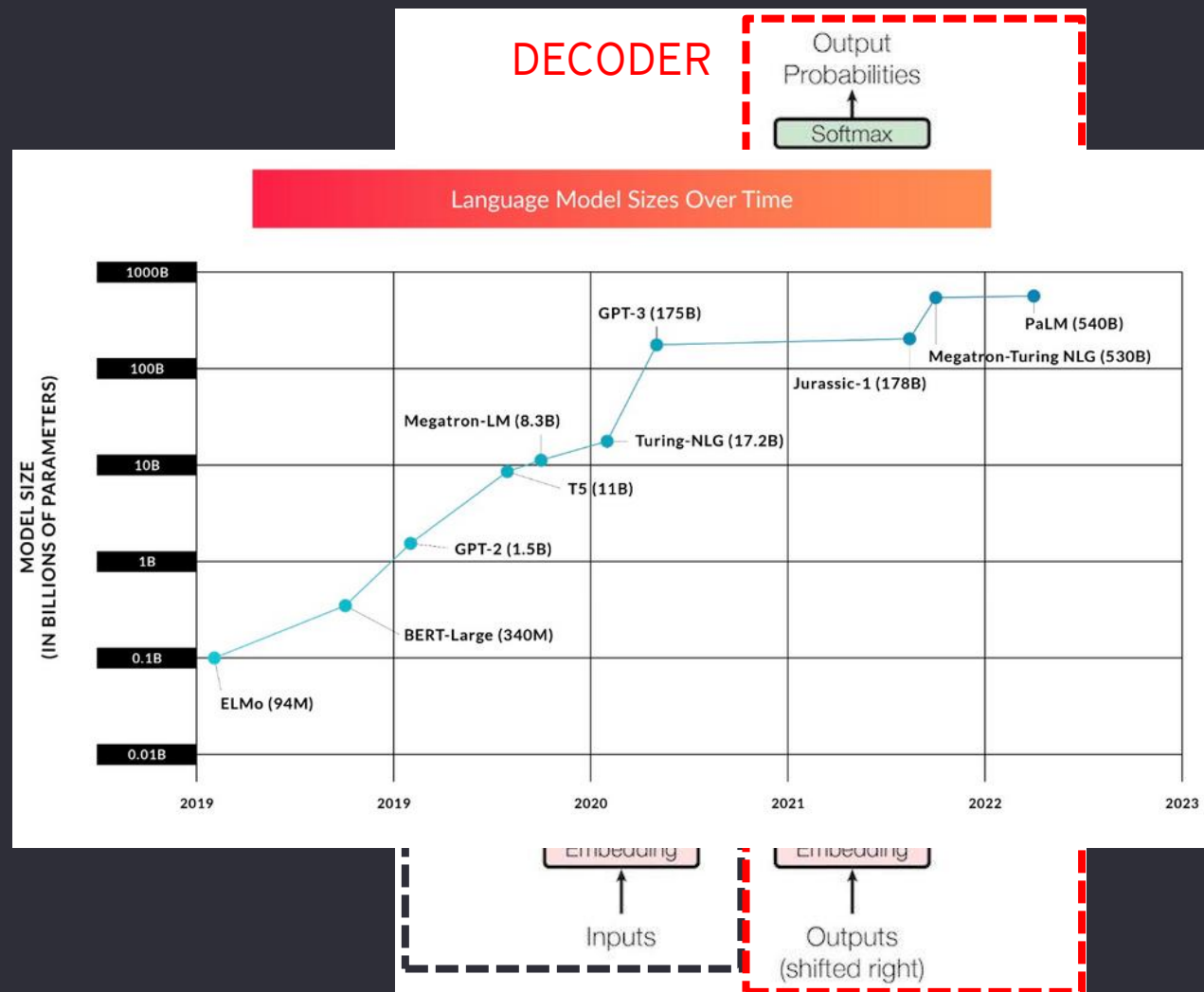


¿Cómo pasamos de LM a LLM?

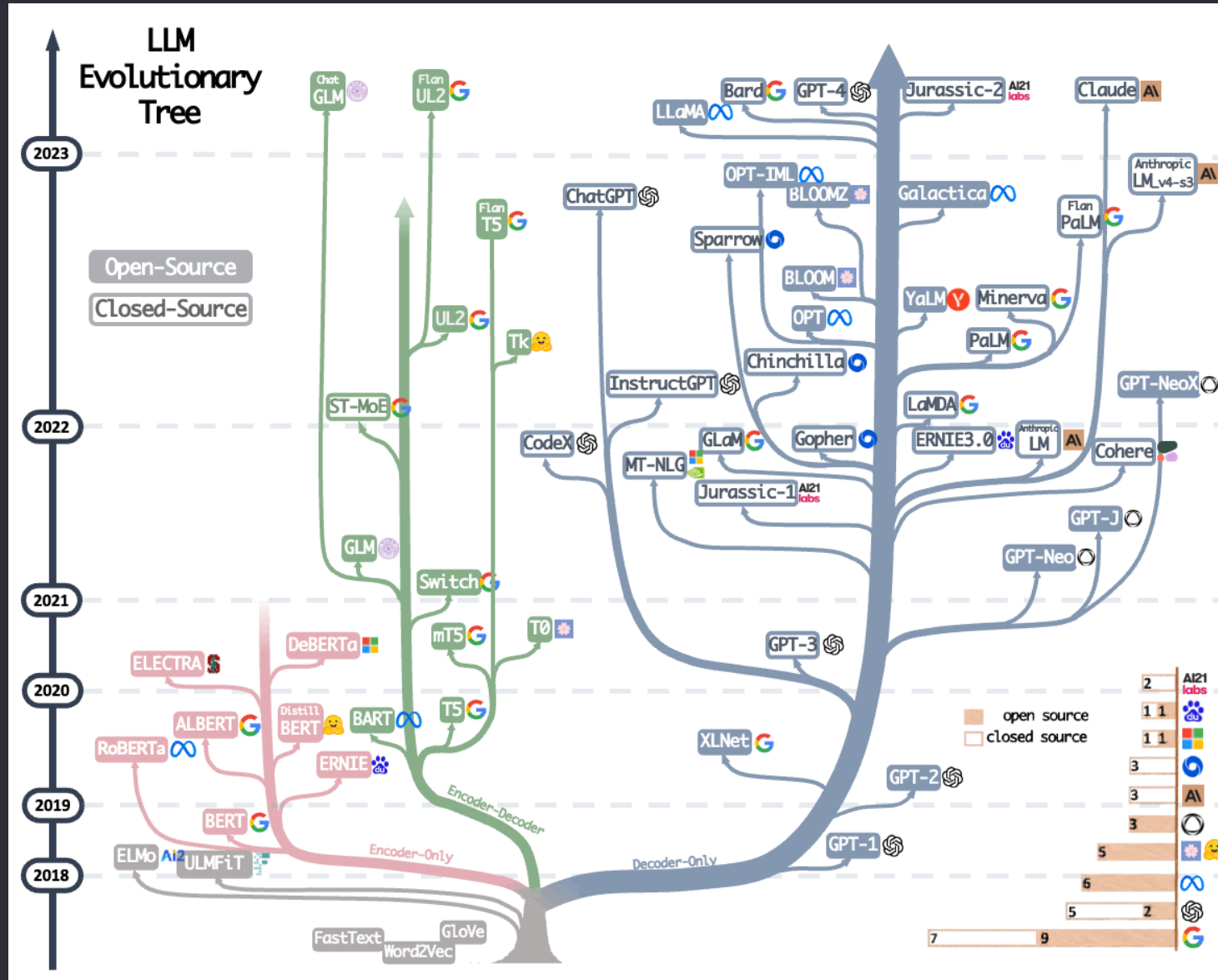
Language Model	Description	"Large"?	Emergence
Bag-of-Words Model	Represents text as a set of unordered words, without considering sequence or context	No	1950s-1960s
N-gram Model	Considers groups of N consecutive words to capture sequence	No	1950s-1960s
Hidden Markov Models (HMMs)	Represents language as a sequence of hidden states and observable outputs	No	1980s-1990s
Recurrent Neural Networks (RNNs)	Processes sequential data by maintaining an internal state, capturing context of previous inputs	No	1990s-2010s
Long Short-Term Memory (LSTM) Networks	Extension of RNNs that captures longer-term dependencies	No	2010s
Transformers	Neural network architecture that processes sequences of variable length using a self-attention mechanism	Yes	2017-Present

Arquitectura Transformers

- Arquitecturas de transformers:
 - Encoder
 - Decoder
 - Ejemplos
 - **BERT** model (Bidirectional Encoder Representations from Transformers)
 - **GPT** models (Generative Pre-trained Transformer)
- Carrera por el desarrollo: aparición de LLMs
- Modelos recientes
 - Gran tamaño (parámetros)
 - **GPT o ELMO** → millones de parámetros
 - **GPT2 o BERT** → cientos de millones de parámetros
 - **PALM o GPT3** → miles de millones de parámetros

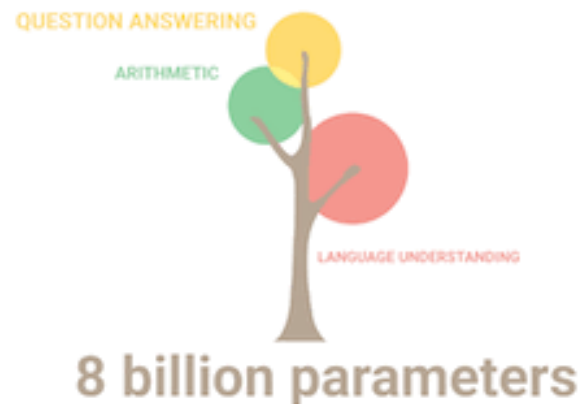


Evolución de los LLM's



Nuevos paradigmas de los LLM

- Los efectos inesperados de escalar los LM
 - Nuevas capacidades para las que no fueron entrenados directamente
 - “Emergence”
- Desafíos del escalamiento
 - Paper “Training Compute-Optimal Large Language Models”, 2023
 - LLM sub-entrenados
 - Parámetros y data deben crecer de manera proporcional

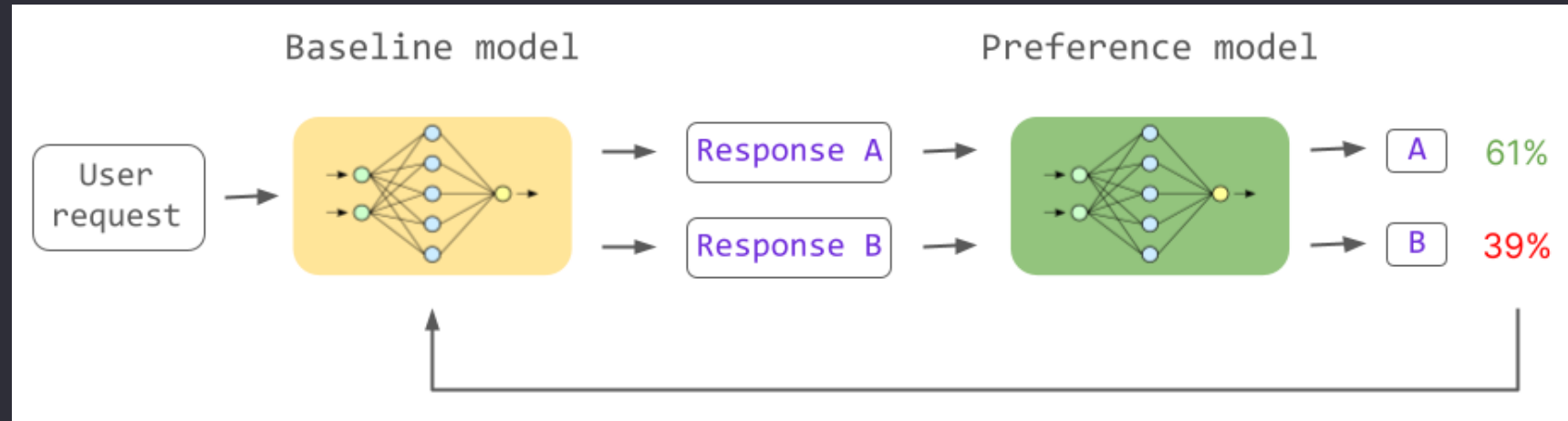


Otros desafíos de los LLM

- Grandes inquietudes: “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” (2021)
- Algunos riesgos inherentes en los modelos.
 - Utilidad o resultado
 - Veracidad
 - Inocuidad

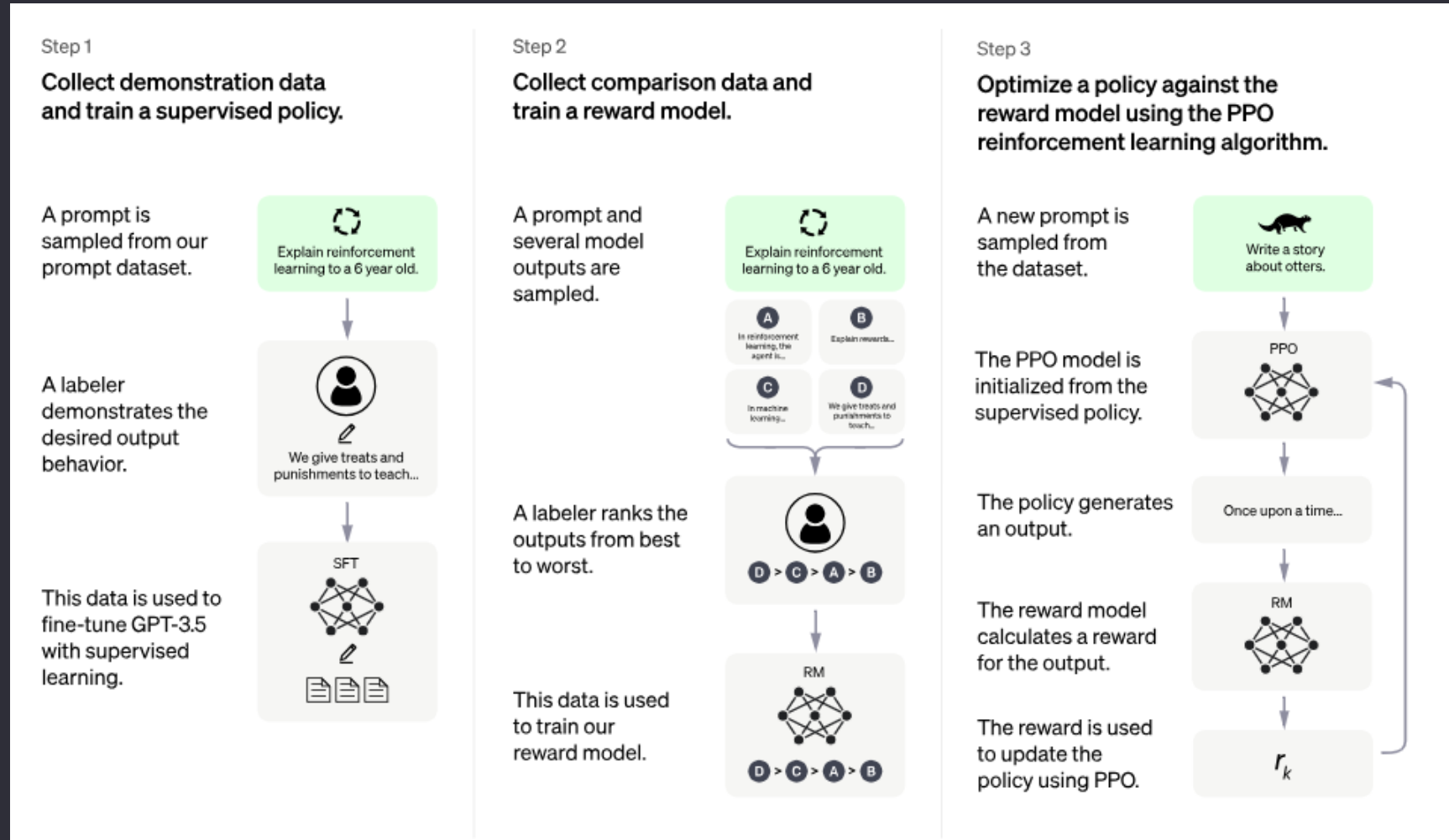
¿Solución?

- RLHF (Reinforcement learning from human feedback)
 - 2 modelos: modelo base + modelo de preferencia.



RFLH en ChatGPT

1. Recolectar data para realizar ajuste fino a GPT 3.5
2. Recolectar data de comparación y entrenar un modelo de recompensas
3. Aprendizaje reforzado a partir de modelo de recompensas.



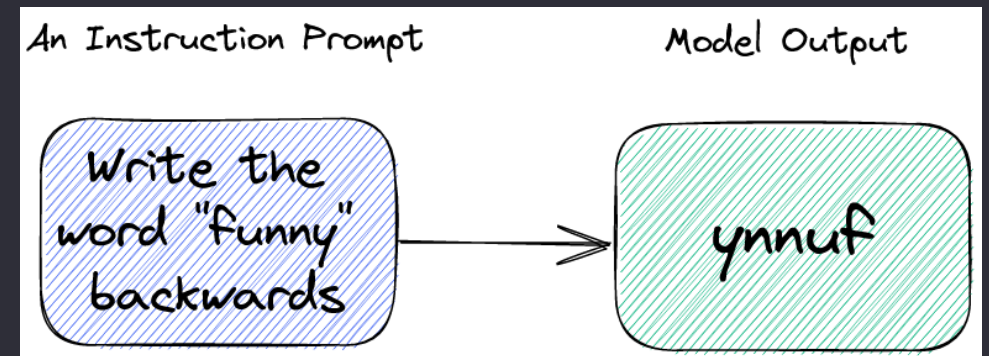
Prompting & Prompt Engineering

Prompting: Proceso por el cual se instruye (dar instrucciones) a una inteligencia artificial para realizar una tarea. Estas instrucciones (el prompt) puede ser tan simple como una instrucción/pregunta o tan complejo como párrafos de texto.

Prompt Engineering: práctica de desarrollar y optimizar prompts para utilizar modelos de lenguajes (LLM's) de manera eficiente para una variedad de usos.

Algunos usos típicos que podemos darle a través de prompts:

- Resumen de texto
- Extracción de información
- Generador de información
- Clasificación de texto
- Generador de código
- Razonamiento



Prompting & Prompt Engineering

Elementos de un prompt*:

- **Instrucción:** tarea que se desea que se realice.
- **Contexto:** involucra información externa o adicional que puede dirigir al modelo a respuestas mejores.
- **Datos de entrada:** entrada o pregunta para la cual se requiere una respuesta.
- **Indicador de salida:** tipo o formato de salida de la respuesta.

Algunos consejos al momento de construir prompts:

- Partir por lo simple y luego ir agregando elementos para apuntar a mejores resultados (versionar)
- Señala claramente que quieres que el modelo haga: 'Resume', 'Clasifica', 'Ordena', 'Traduce'. Se puede probar que funciona mejor en cada caso. Trate de decir que debe hacer y no lo que no debe hacer.
- Mientras más descriptivo y detallado, mejor será el resultado. Pero estos detalles deben ser relevantes y no se debe caer en descripciones imprecisas. (Comunicación efectiva → mientras más directo, más efectivo).
- Es buena idea proporcionar ejemplos pero hay que tener cuidado con la longitud del prompt.

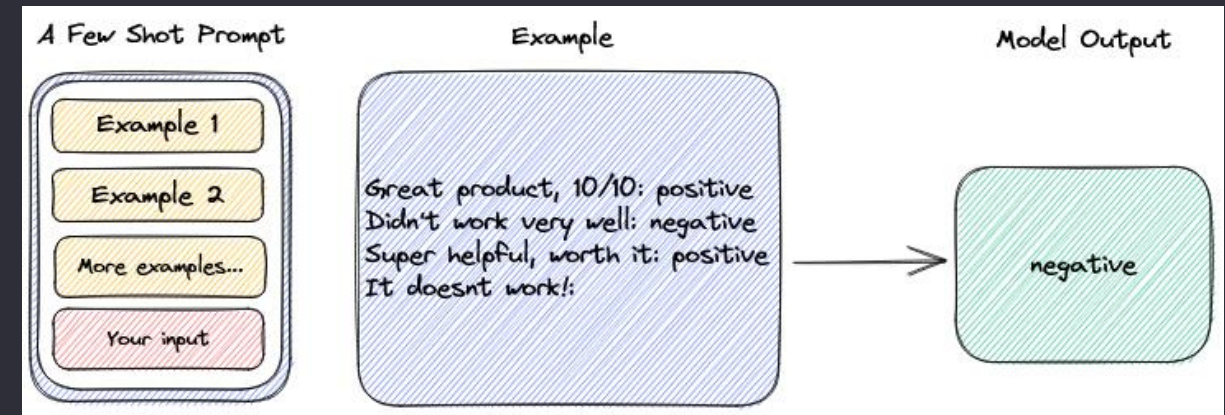
* No todos estos elementos son necesarios para un prompt

Técnicas de Prompting

Few shot prompting*

Técnica que implica mostrarle al modelo algunos ejemplos sobre lo que se quiere hacer.

- Un caso importante para usar esta técnica es cuando necesitamos que el resultado este estructurado de una manera específica difícil de describir



Variantes de shot prompting

Existen otros tipos de shot-prompting que dependerán solamente de la cantidad de ejemplos que se le muestran al modelo.

Variantes:

- **0 shot prompting:** no se muestran ejemplos
- **1 shot prompting:** un ejemplo es proporcionado
- **few shot prompting:** 2 o más ejemplos son dados.

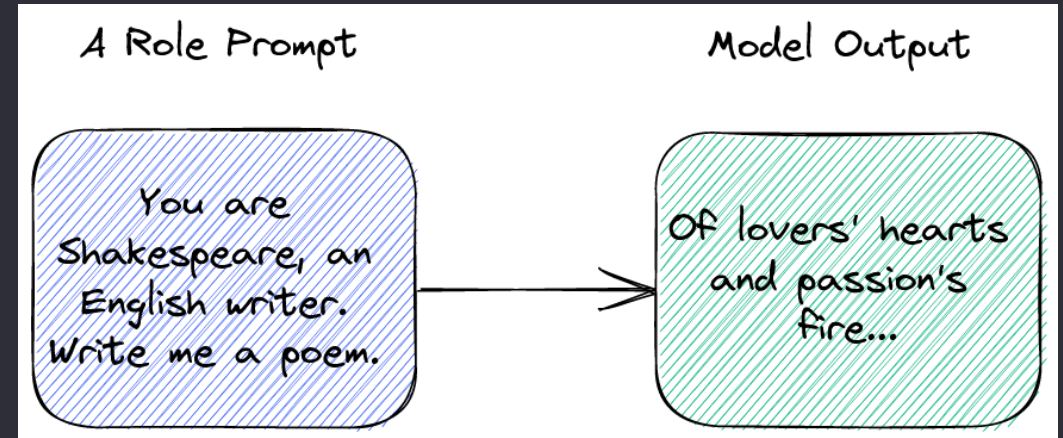
* Si bien funciona para algunas tareas, se debe tener cuidado al utilizar con tareas complejas.

Técnicas de Prompting

Role Prompting

Técnica que le asigna un rol al modelo. Cuando hacemos esto le estamos dando un contexto para entender mejor la pregunta.

** El modelo contiene respuestas que todas las personas podrían dar. Como no todas las personas tienen el mismo conocimiento sobre un tópico, retornará una respuesta promedio. Al asignarle un rol, estamos "podando" aquellas respuestas que no tienen un conocimiento acabado.*



Técnicas de Prompting

Cadena de pensamiento (*Chain of Thought*)

El prompt por cadena de pensamientos (CoT) es una técnica que motiva al modelo a explicar su razonamiento en pasos intermedios, lo que permite capacidades de razonamiento más complejas. Puedes combinarlo con prompts de ejemplo (few-shot) para obtener mejores resultados en tareas más complejas que requieren razonamiento antes de responder.

- **Limitaciones**

"CoT only yields performance gains when used with models of ~100B parameters".

- Modelos más pequeños escriben cadenas de pensamientos con falta de lógica, empeorando el resultado final.

(a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

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:

(Output) The answer is 8. ✗

(b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

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:

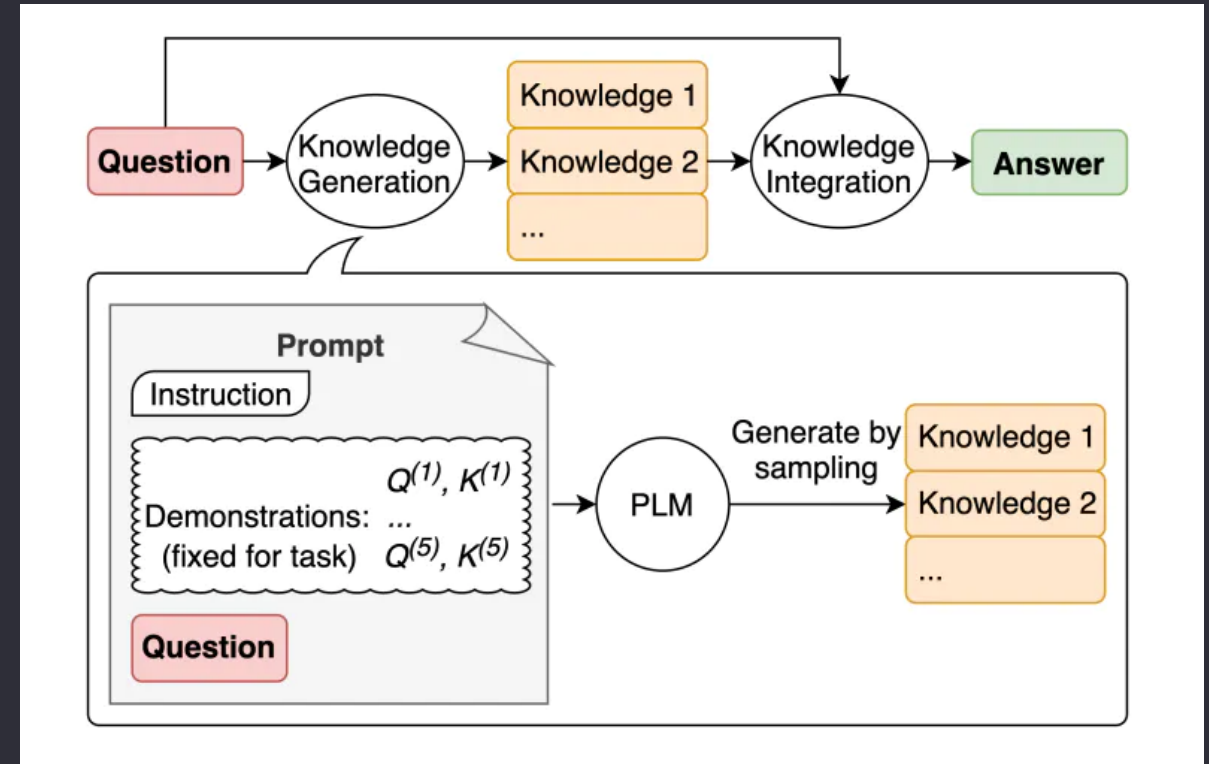
(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are $16 / 2 = 8$ golf balls. Half of the golf balls are blue. So there are $8 / 2 = 4$ blue golf balls. The answer is 4. ✓

Técnicas de Prompting (otros)

Prompt de conocimiento generado

Técnica que incluye la capacidad de incorporar conocimiento o información para ayudar al modelo a hacer predicciones más precisas.

Usando una idea similar, ¿puede el modelo también ser utilizado para generar conocimiento antes de hacer una predicción?



Técnicas de Prompting

Prompt de Estímulo direccional

Es una técnica reciente donde se propone dar una guía para mejorar la tarea introduciendo ciertos estímulos direccionales

Originalmente se propuso crear estos estímulos mediante un LM ajustable, pero estos tips o hints también puede ser dados por el usuario para obtener los resultados deseados.

Input text

Article: (CNN) For the first time in eight years, a TV legend returned to doing what he does best. Contestants told to "come on down!" on the April 1 edition of "The Price Is Right" encountered not host Drew Carey but another familiar face in charge of the proceedings. Instead, there was Bob Barker, who hosted the TV game show for 35 years before stepping down in 2007. Looking spry at 91, Barker handled the first price-guessing game of the show, the classic "Lucky Seven," before turning hosting duties over to Carey, who finished up. Despite being away from the show for most of the past eight years, Barker didn't seem to miss a beat.

Reference

Bob Barker returned to host "The Price Is Right" on Wednesday . Barker, 91, had retired as host in 2007 .

Directional Stimulus Prompting

Q: Summarize the above article briefly in 2-3 sentences **based on the hint**.
Hint: Bob Barker; TV; April 1; "The Price Is Right"; 2007; 91.

Model output

A: On April 1, Bob Barker returned to the TV show "The Price Is Right" after eight years . Looking spry at 91, Barker handled the first price-guessing game . Barker stepped down as host in 2007, but didn't seem to miss a beat .

ROUGE-1: **48.39** 👍

Standard Prompting

Q: Summarize the above article briefly in 2-3 sentences.

Model output

A: Bob Barker, the former host of \"The Price Is Right\", returned to the show for the first time in 8 years. He hosted the first game of the show, and then handed the hosting duties over to Drew Carey.

ROUGE-1: **34.48** 👎

Algunas dificultades de LLMs

- **Cita de fuentes:** algunos LLMs no puede citar precisamente fuentes de información. (ya sea porque no tienen acceso a internet o no recuerdan exactamente de donde vino la información)
 - **Sesgos:** en general las LLMs están sesgadas a generar respuestas estereotípicas. Existe la probabilidad de que se generen respuestas que pueden considerarse reprochables. Se debe tener cuidado con respecto a aplicaciones de cara a clientes finales.
 - **Alucinaciones:** LLMs también puede generar aseveraciones falsas frente a una pregunta de la cual no conocen su respuesta.
 - **Matemáticas:** Puede que existan dificultades en resolver algunos problemas matemáticos. Hay algunos LLMs que pueden ser parchados o ocupando otras APIs para generar la respuesta correcta.
 - **Prompt Hacking:** Existe la posibilidad de engañar a la LLM y generar respuestas que se quiera o literalmente evadir alguna política establecida en su configuración.
-

Para tener en consideración

- Al momento de construir un producto se debe tener en consideración que modelo se está utilizando.
Frente a un mismo prompt distintos modelos generaran distintas respuestas.
 - Para poder tener resultados robustos, es necesario considerar los siguientes pasos:
 - **Paso 1:** Investiga
 - Modelos que se quieren usar
 - Prompt que se quiere utilizar (ver recursos extra)
 - **Paso 2:** Experimentar e iterar
 - Utiliza un esquema de refinación, testea las respuestas generadas y adapta tus prompts de acuerdo al objetivo final.
 - Prueba agregar contexto para mejorar los resultados.
 - Pruebas distintas técnicas de prompting para ver cual es la que mejor se adapta a tus requerimientos
-

Algunos recursos extra



1. <https://www.deeplearning.ai/short-courses/>
 2. <https://www.promptingguide.ai/es>
 3. <https://learnprompting.org/>
 4. <https://platform.openai.com/docs/introduction>
 5. <https://github.com/openai/openai-cookbook>
 6. <http://pretrain.nlpedia.ai/>
 7. <https://github.com/dair-ai/Prompt-Engineering-Guide>
 8. <https://docs.cohere.com/docs/prompt-engineering>
-

Para practicar / experimentar:



1. <https://www.perplexity.ai/>

2. <https://bard.google.com/?hl=en>

3. <https://openai.com/chatgpt>



Preguntas o Consultas

Avisos y fechas importantes

Fechas de sesiones

- ~~Sesión 1: martes 12 de septiembre - "Design Thinking"~~
 - ~~Sesión 2: miércoles 13 de septiembre - "LLM & Prompting"~~
 - Sesión 3: jueves 14 de septiembre - "APIS de OpenAI"
 - Sesión 4: martes 26 de septiembre - "Langchain"
 - Sesión 5: miércoles 27 de septiembre - "Casos de Uso / Wireframing"
 - Sesión 6: jueves 28 de septiembre 2023 - "Pitch Class y Aclaratorias generales"
- (Sesión 6 es OBLIGATORIA, deben participar a lo menos 2 integrantes del equipo)**





HA
CKA
TÓN

XI
20
23

SUMMIT
PAÍS
DIGITAL

FUNDACIÓN
país
digital

By:



Colabora:

