Cursos Extraordinarios

Verano 2024

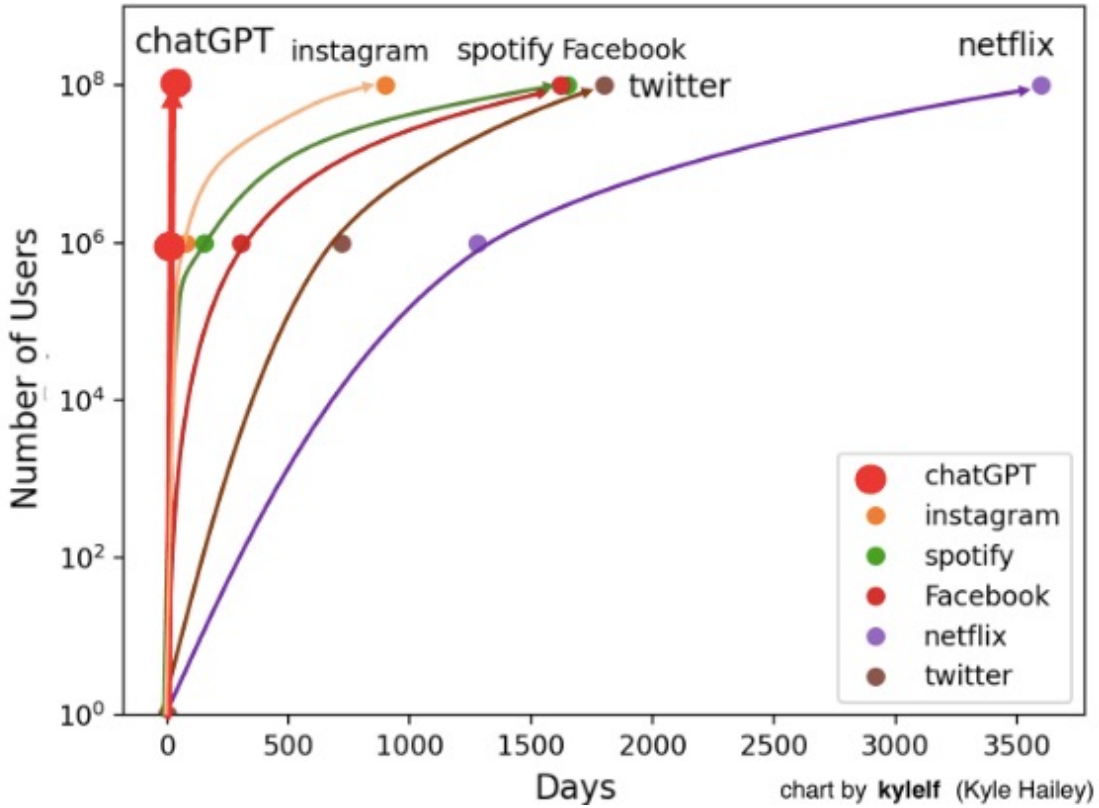
“Inteligencia Artificial y Grandes Modelos de Lenguaje: Funcionamiento, Componentes Clave y Aplicaciones”

Zaragoza, del 3 al 5 de julio

Grandes Modelos de Lenguaje

Introducción

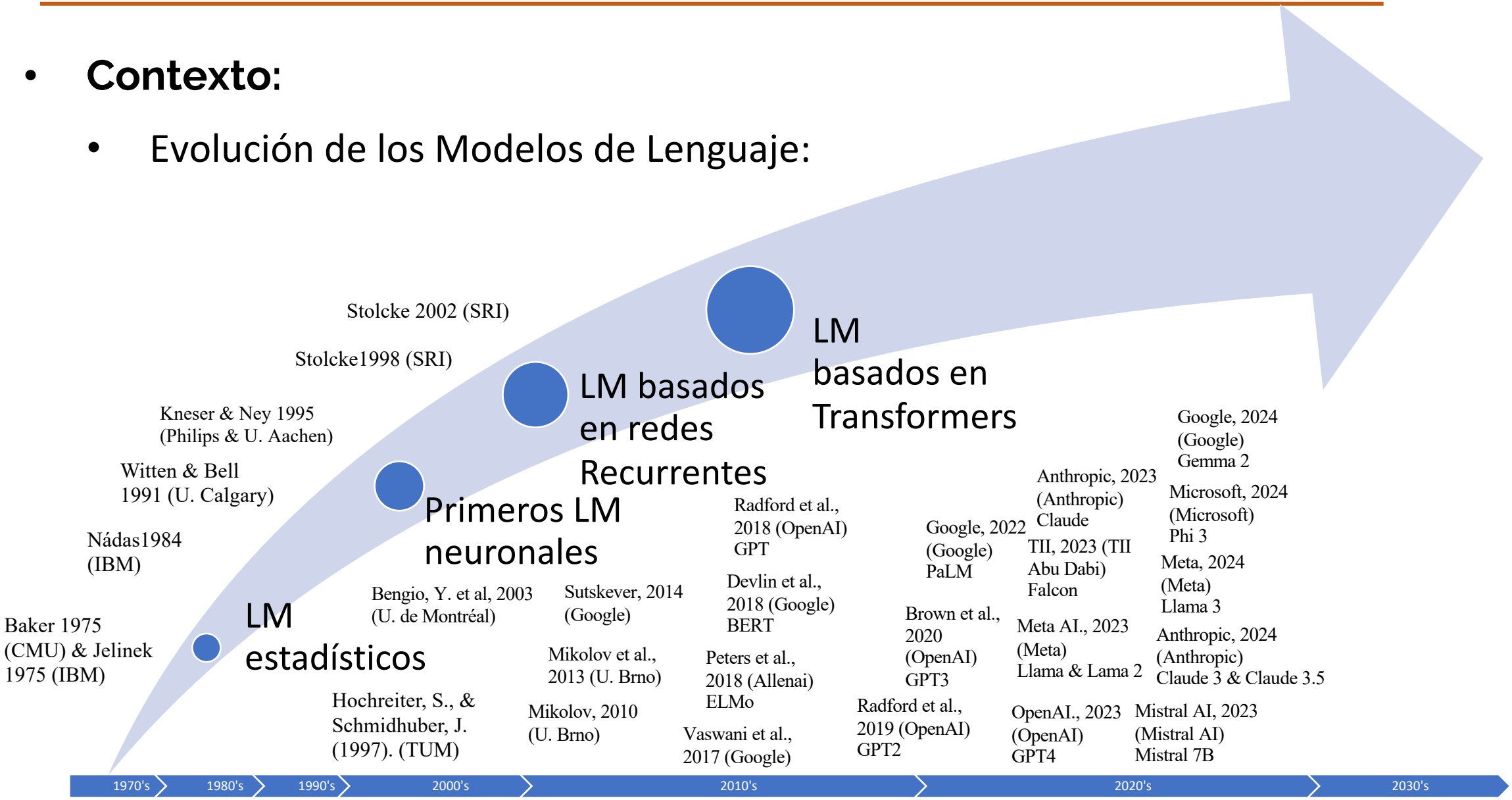
- **Contexto:**
  - ChatGPT fue lanzado a finales de noviembre 2022.
  - Repercusión mediática sin precedentes:



Evolución del número de usuarios en función de los días transcurridos desde su lanzamiento para diversos productos tecnológicos

Introducción

- **Contexto:**
  - Evolución de los Modelos de Lenguaje:



Large Language Models

- **Transformer-Based:**

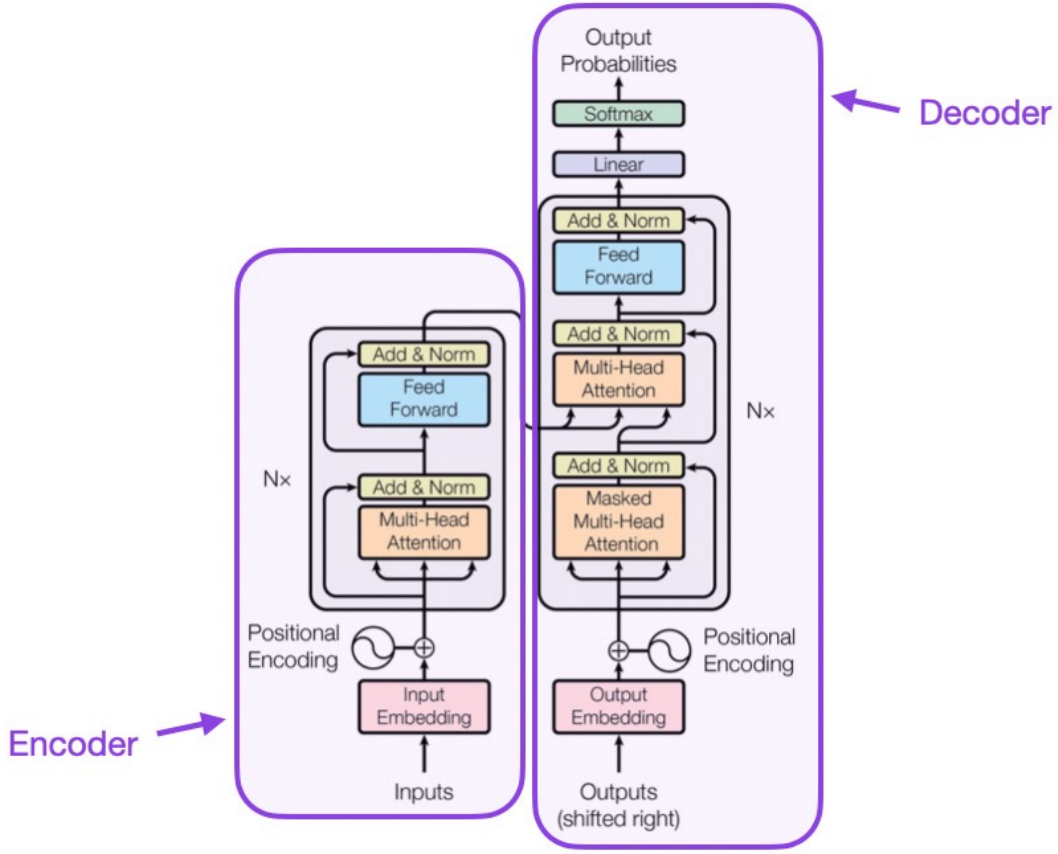


Figure 1: The Transformer - model architecture.

- **Encoder Only:**
- **Decoder Only:**
- **Encoder-Decoder :**

Large Language Models

- **Transformer-Based:**

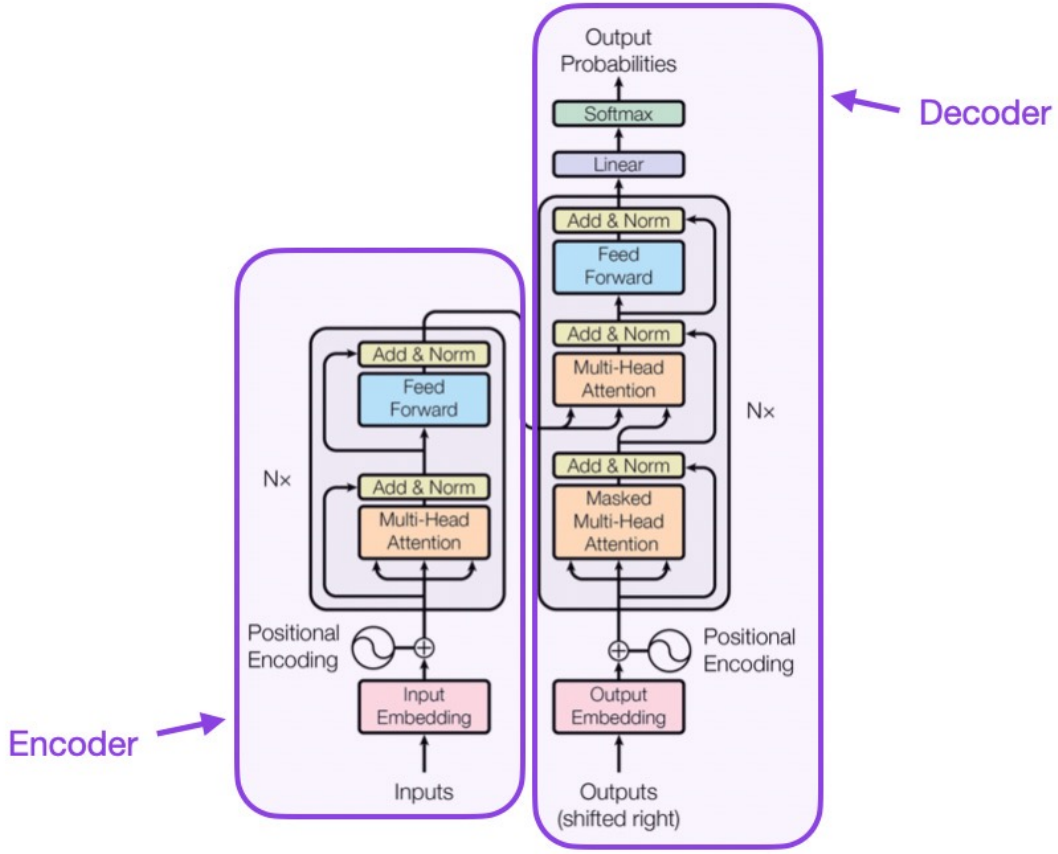
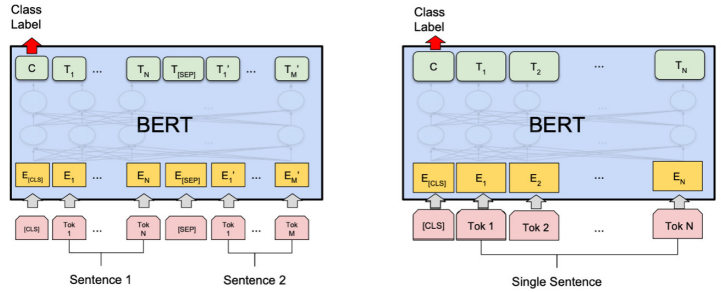


Figure 1: The Transformer - model architecture.

- **Encoder Only:**
  - **BERT, RoBERTa**

Encoder-style BERT model for predictive modeling tasks



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

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

- **encoder-style transformers for predictive modeling tasks such as text classification**

***BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (2018)** by Devlin, Chang, Lee, and Toutanova https://arxiv.org/abs/1810.04805

RoBERTa: A Robustly Optimized BERT Pretraining Approach by Liu, Ott, et al. (2019) https://arxiv.org/pdf/1907.11692

Large Language Models

- **Transformer-Based:**

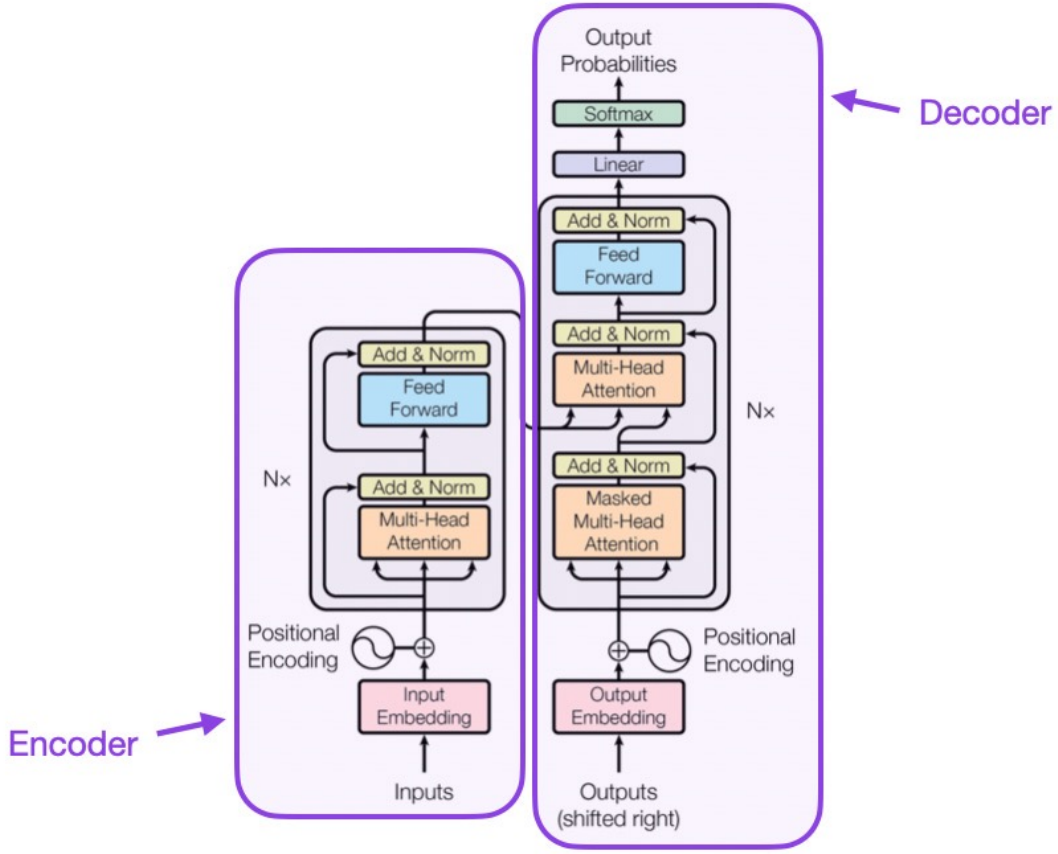
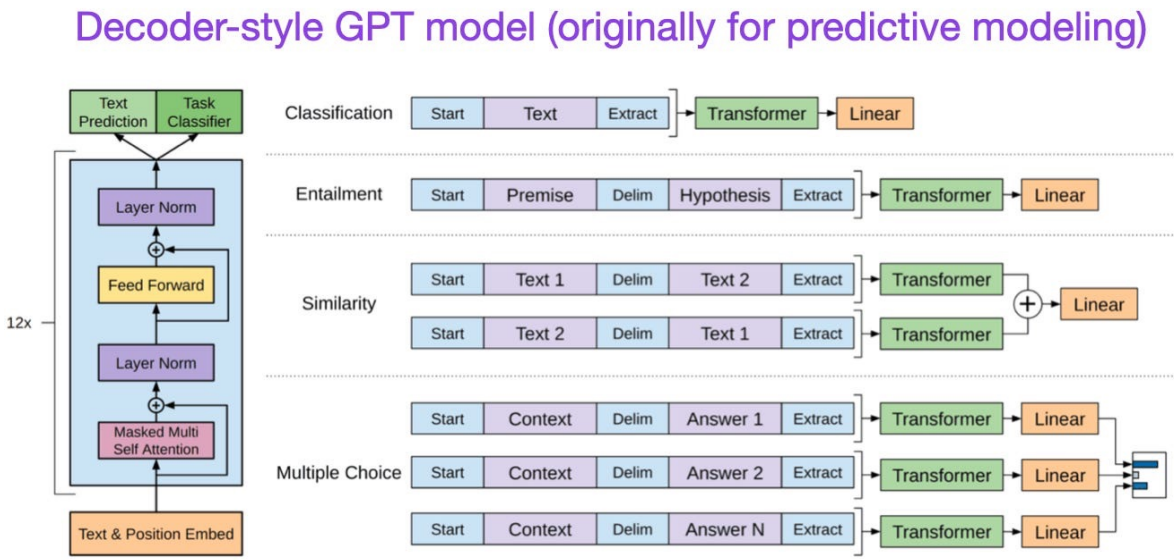


Figure 1: The Transformer - model architecture.

- **Decoder Only:**
  - **GPT**



- **BERT is bidirectional with masked language model pretraining objective, GPT is unidirectional, autoregressive model**

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

Large Language Models

- **Transformer-Based:**

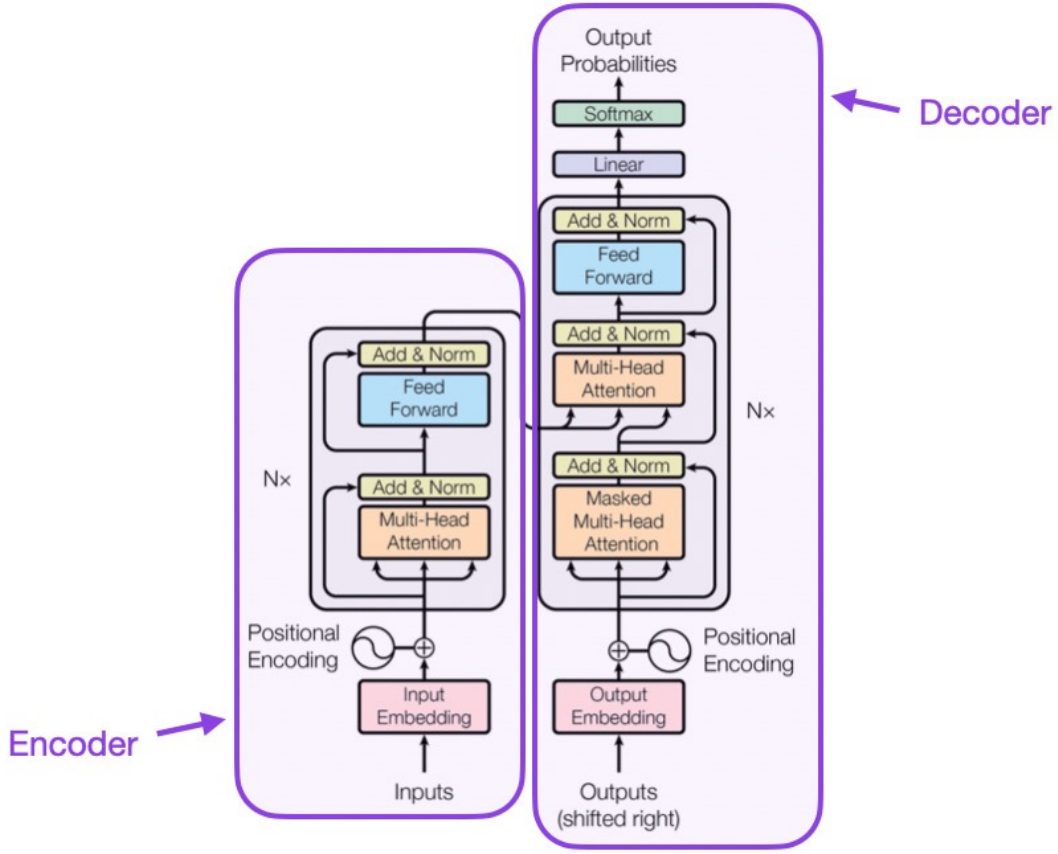
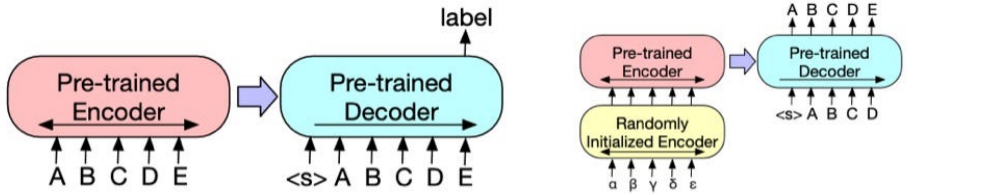


Figure 1: The Transformer - model architecture.

- **Encoder-Decoder :**
  - **BART**

BART combines encoder and decoder parts



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

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

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

Large Language Models

- **Transformer-Based:**

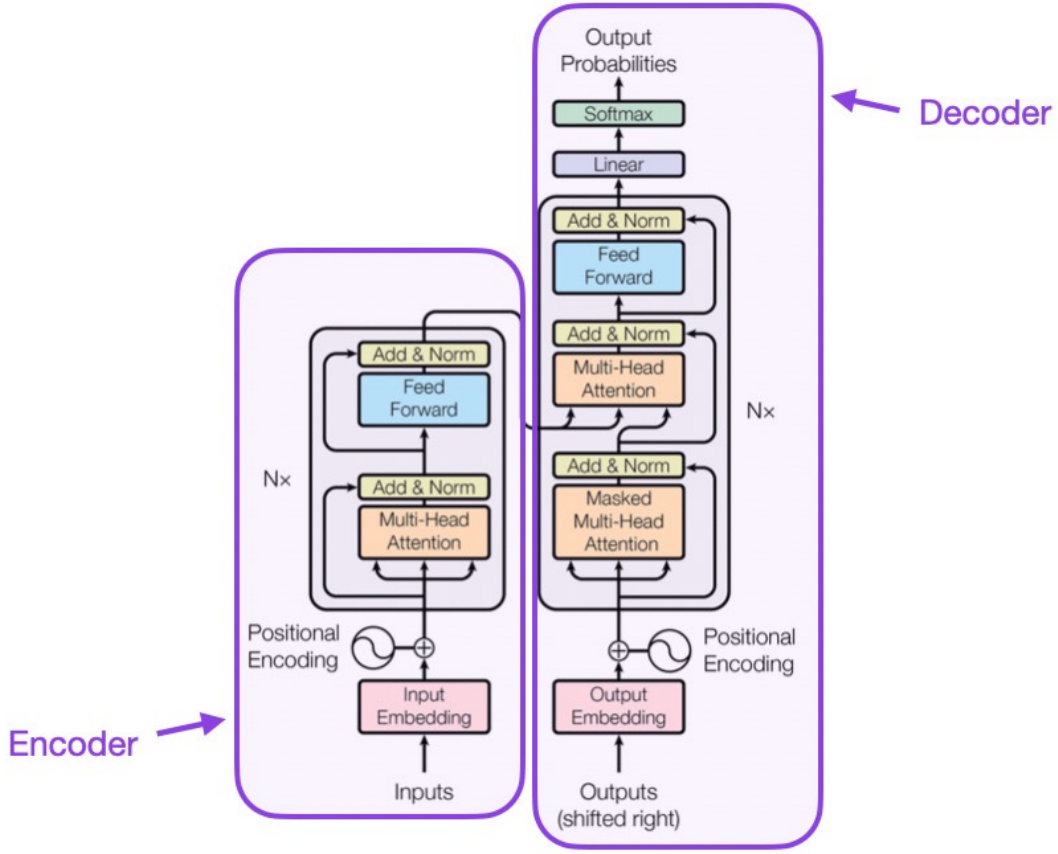


Figure 1: The Transformer - model architecture.

- **Encoder Only:**
  - Entrenamiento basado en la predicción de palabras enmascaradas de modo discriminativo
- **Decoder Only:**
  - Entrenamiento generativo basado en la predicción de la siguiente palabra (autorregresivo)
- **Encoder-Decoder :**
  - Entrenamiento basado en la predicción de palabras enmascaradas de modo discriminativo

Large Language Models

- **Transformer-Based:**

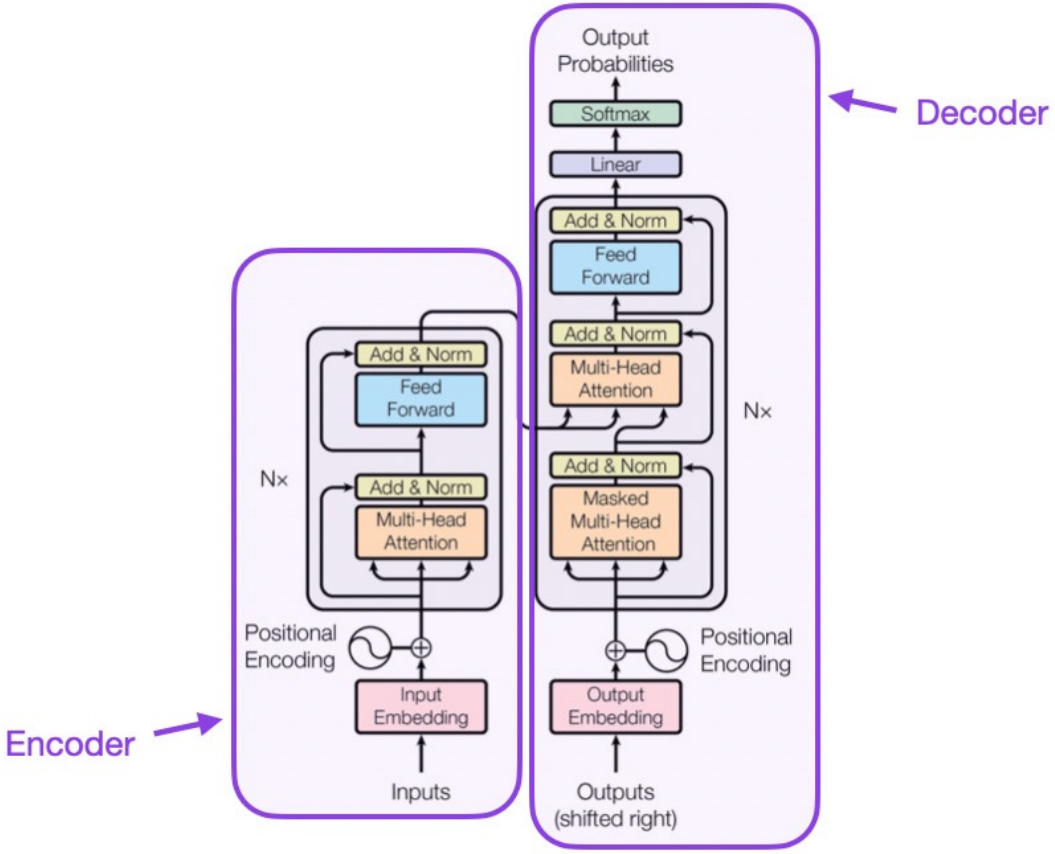
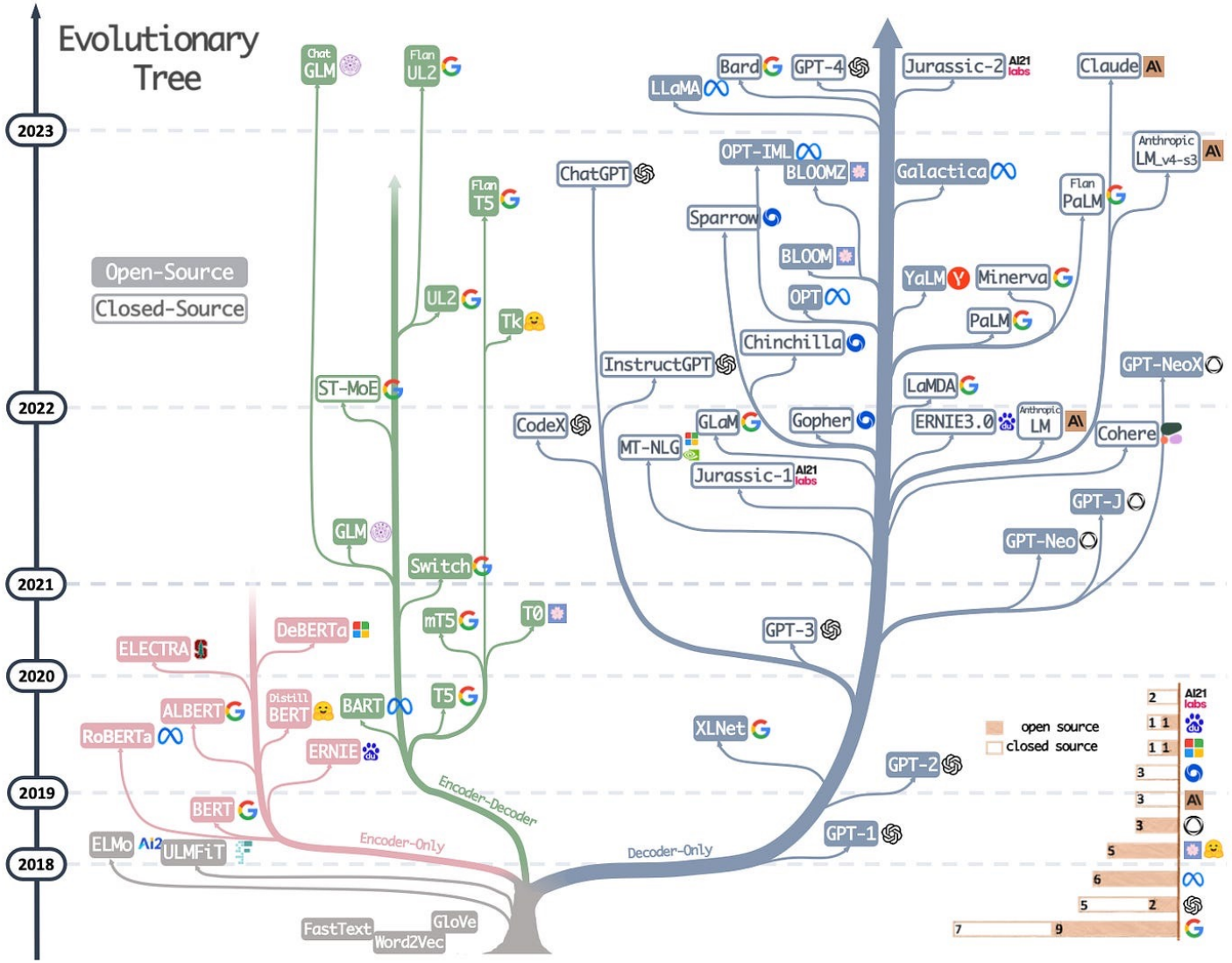


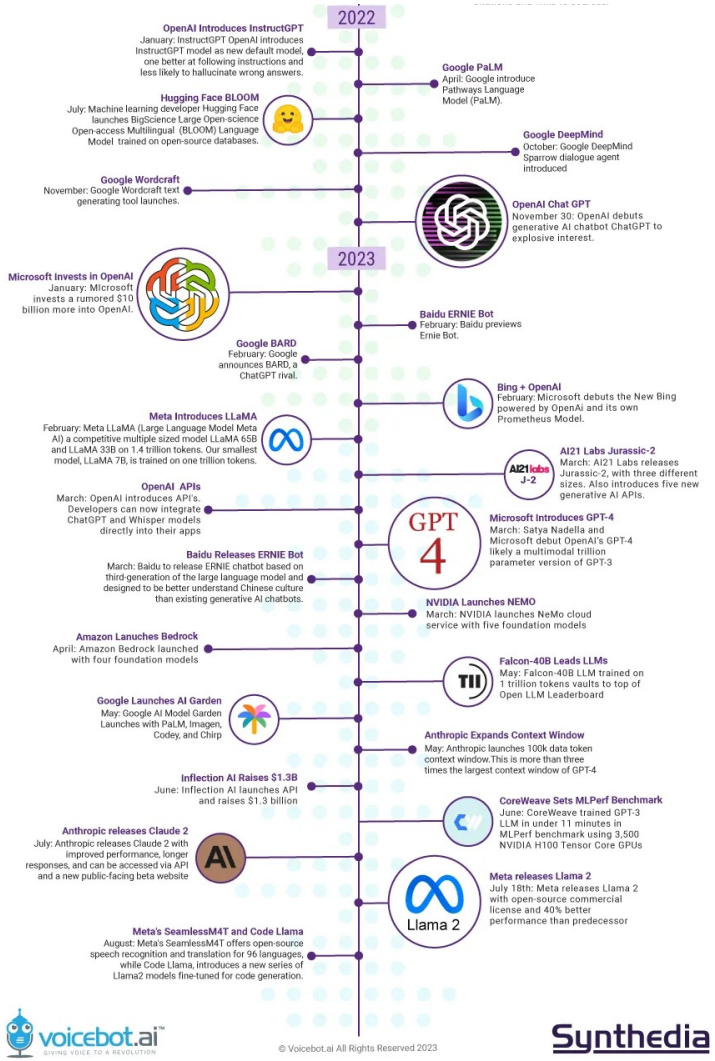
Figure 1: The Transformer - model architecture.

- **Encoder-Decoder :**
  - Entrenamiento basado en la predicción de palabras enmascaradas de modo discriminativo
  - Desde el principio y sin fine-tuning ya da resultados razonables en determinadas tareas: Q&A, writting, …
  - Información no fiable, alucinaciones, …

Large Language Models

- **Evolución Temporal:**

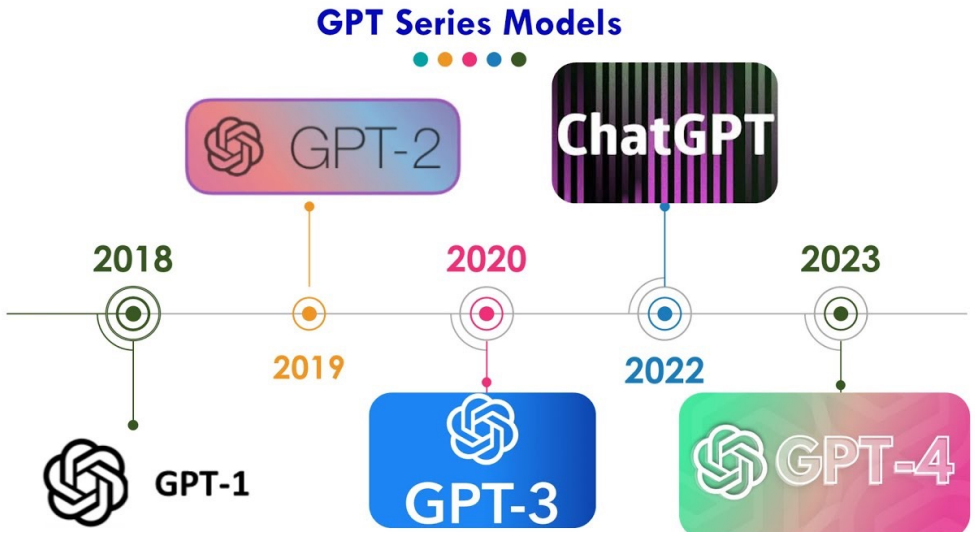




Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond
Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, Xia Hu

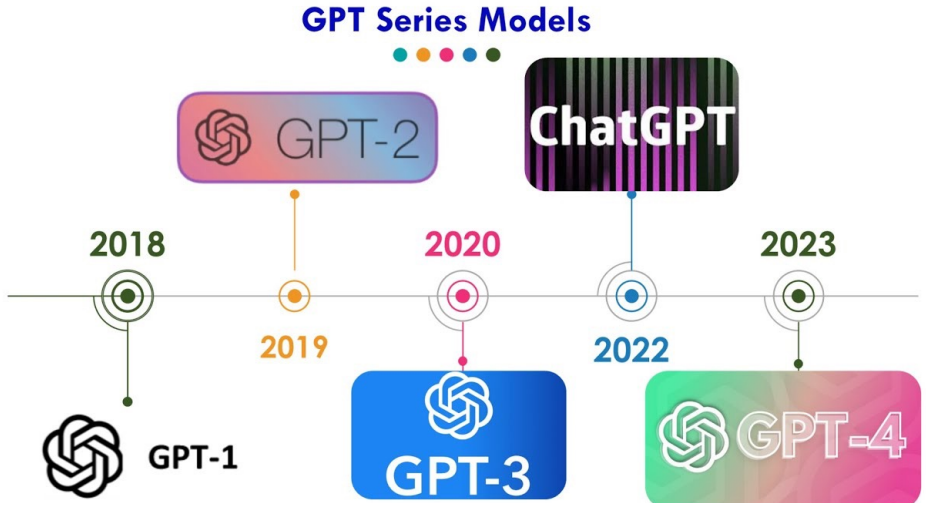
Large Language Models

- **Modelos generativos, GPT:**
  - Primera versión GPT 2018
  - Intento de resolver varias tareas de NLU, mediante modelo preentrenado + finetuning
  - 117 Millones de parámetros
  - Pre “leyes de escala de los LLM”



Large Language Models

- **Modelos generativos, GPT:**
  - Segunda versión 2019: GPT-2
  - 1500 Millones de parámetros
  - Multitarea mediante entrenamiento no supervisado sin finetuning
  - Salida condicionada a la tarea:
    - Salida dada la entrada, pero condicionada a la tarea
    - La tarea viene descrita en lenguaje natural con formato textual
  - Prestaciones todavía no superiores a modelos específicamente diseñados para las tareas concretas: Q&A, comprensión, resumen, …
  - Primeras muestras de capacidades zero-shot o few-shot



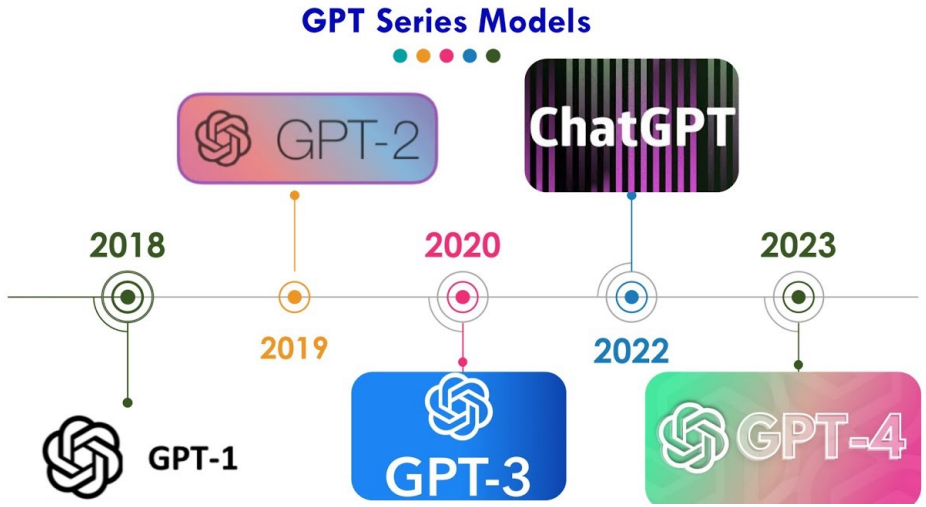
Large Language Models

- **Modelos generativos, GPT:**
  - Segunda versión 2019: GPT-2

System Prompt (human-written)	*In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.*
Model Completion (machine-written, 10 tries)	The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science. Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved. Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow. Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez. Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

Large Language Models

- **Modelos generativos, GPT:**
  - Tercera versión 2020: GPT- 3
  - 175000 Millones de parámetros
  - Capacidad para desempeñar tareas (a partir de unos pocos ejemplos) para la que no fue entrenado:
    “task-agnostic” few-shot Learning
  - Competitivo frente a modelos específicamente entrenados para las tareas
  - Generación de textos indistinguibles de los textos humanos
  - Capacidad “in-context Learning” que permite tareas ”zero-shot” o “few-shot”:
    El modelo infiere del contexto lo que debe hacer y responde en consecuencia



Large Language Models

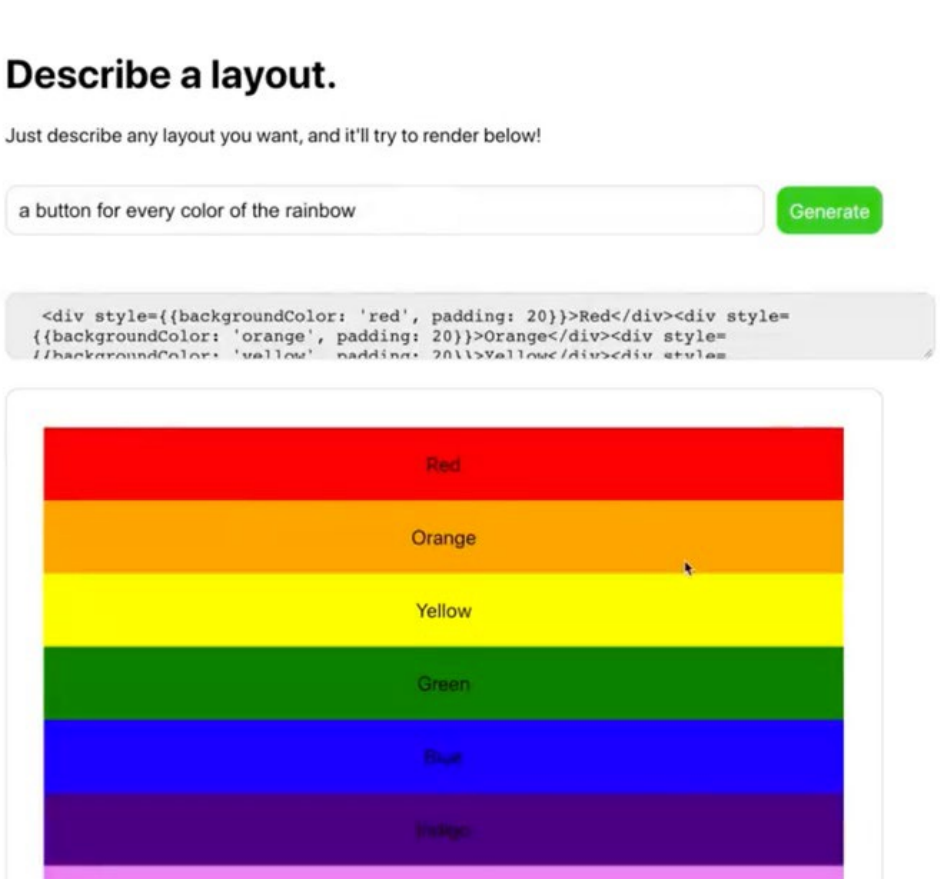
- **Modelos generativos, GPT:**
  - Tercera versión 2020: GPT- 3

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 *bold* 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 *his* financial support. Question: In the passage above, what does the pronoun "*his*" refer to? Answer:
Target Completion →	mr. moncrieff

”zero-shot”

Large Language Models

- **Modelos generativos, GPT:**
  - Tercera versión 2020: GPT- 3



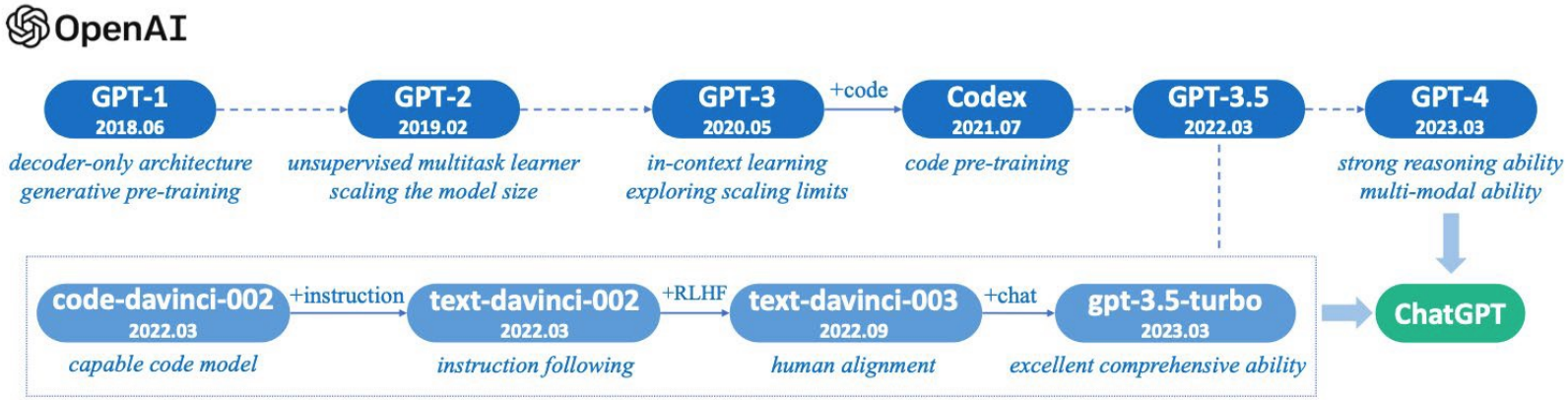
”few-shot”: Generación de layouts con JSX

Large Language Models

- **Modelos generativos, GPT:**
  - **Refinamientos sobre GPT 3 sobre datos específicos en dos líneas**:
    - Modelo para generación de código: Codex (2021)
      - Fine-tuned con repositorios de código (Repositorios públicos de GitHub).
      - Bueno programando y Bueno también resolviendo problemas matemáticos:
        - GPT-3.5 (code-davinci-002):
          - Bueno en tareas que requieren de capacidades “chain-of-thought”.
          - Se fuerza que el modelo, al generar su respuesta, explique el “razonamiento” que lleva a la generación de dicha respuesta.
    - Modelo ajustado a los estilos y preferencias humanas (human alignment): InstructGPT (2022)
      - Reinforcement Learning: Aprender a partir de respuestas o “preferencias” de humanos. (RLHF)
      - Desarrollo de modelos que ”predicen” las preferencias humanas: *reward model*
      - Protección frente a la generación de contenido “tóxico”

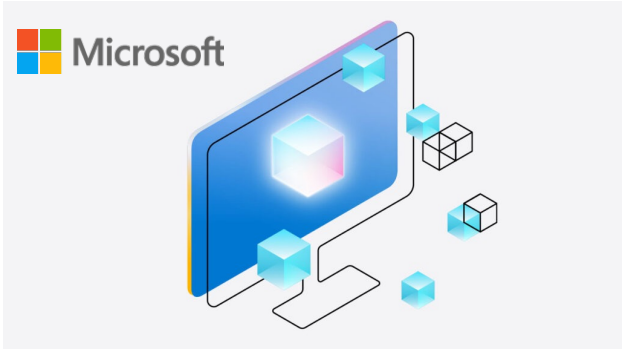
Large Language Models

- **Modelos generativos, GPT:**
  - Noviembre de 2022, siguiendo la línea de InstructGPT: ChatGPT
    - Sistema conversacional entrenado de modo similar a InstructGPT especializado en la tarea de diálogo
      - Conversaciones generadas por humanos en las que ambos actuaban como usuario y agente entre los datos de fine-tuning.
      - Mayor capacidad de comunicación con humanos
      - Amplio conocimiento intrínseco
      - Cierta capacidad de razonamiento para problemas matemáticos.



Large Language Models

- **Modelos Abiertos (más recientes):**
  - **Mistral (Mistral AI_), (Septiembre, 2023):**
    - 7B y 7.3B.
    - Contexto 8K
  - **LLAMA 3 (META), (Abril 2024)**:
    - Modelo preentrenado y instruction-fine-tuned
    - 8B y 70B, Anuncian que habrá un modelo 400B
    - Contexto 8k -> hasta 1m con llama3 Gradient (Rotary Position Embedding, RoPE)
  - **Phi 3 (Microsoft), (Mayo 2024)**
    - Modelo multimodal (lenguaje + visión): Phi-3-visión 4.2B
    - Phi-3-mini (3.8B), Phi-3-small (7B) y Phi-3-médium (14B).
      - Contexto 4K y 128K
  - **Gemma 2 (Google), (27 Junio 2024)**
    - 9B y 27B
    - Contexto 8.2K



Google
Gemma 2

MISTRAL
AI_

Large Language Models

- **Modelos Abiertos:**

			Gemma 2		Llama 3		Grok-1
	BENCHMARK	METRIC	9B	27B	8B	70B	314B
General	MMLU	5-shot, top-1	71.3	75.2	66.6	79.5	73.0
Reasoning	BBH	3-shot, CoT	68.2	74.9	61.1	81.3	–
	HellaSwag	10-shot	81.9	86.4	82	–	–
Math	GSM8K	5-shot, maj@1	68.6	74.0	45.7	–	62.9 (8-shot)
	MATH	4-shot	36.6	42.3	–	–	23.9
Code	HumanEval	pass@1	40.2	51.8	–	–	63.2 (0-shot)

MMLU (Massive Multitask Language Understanding)

Large Language Models

- **Modelos Abiertos:**