



ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

EDUCACIÓN
PROFESIONAL

Fundamentos Machine & Deep Learning

Diplomado Big Data
y Analítica de Datos 2022-2023

Profesor:

Rodrigo Sandoval U.





10. Modelos Generadores Avanzados



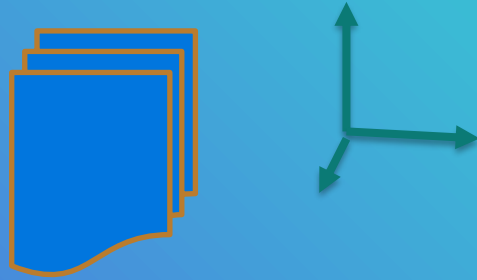
INTRODUCCIÓN

Comenzando con el paper “Attention is all you need”, de 2017, se abre una puerta enorme, impulsada por diferentes organizaciones abiertas, para el desarrollo de modelos avanzados, que combinan varios de los conceptos de Deep Learning (redes neuronales, autoencoders, modelos secuenciales, etc.) para producir modelos capaces de asociar diferentes tipos de información, en la forma de secuencias, para generar resultados derivados de la predicción de lo que viene en la secuencia y aprovecharlos en diferentes contextos.

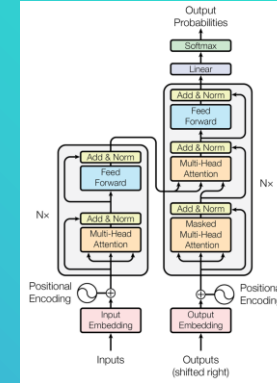
En particular, se ha visto la proliferación de modelos como GPT, BERT, DALL-E y Stable Diffusion, todos usando técnicas relativamente similares, para resolver problemas de diferente índole.

La Brevísima Historia de los LLMs

LLM: Large Language Model



Vectorización de Palabras
Word Embedding
W2V GLoVe FastText



“Attention is all you need”

LLMs

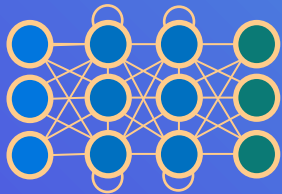
ChatGPT

Dall-E

Stable LLaMA

Diffusion

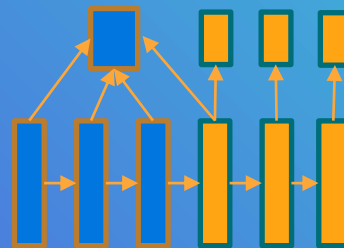
2005



Modelos Secuenciales
RNN LSTM GRU

2010

2015



Mecanismo
atención

2020

Google
BERT

OpenAI
GPT

2022

2023



Attention is all you Need - 2017

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an **encoder and a decoder**. The best performing models also connect the encoder and decoder through an **attention mechanism**. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely.

Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train.

Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

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

Lukasz Kaiser*
Google Brain
lukaszkaier@google.com

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

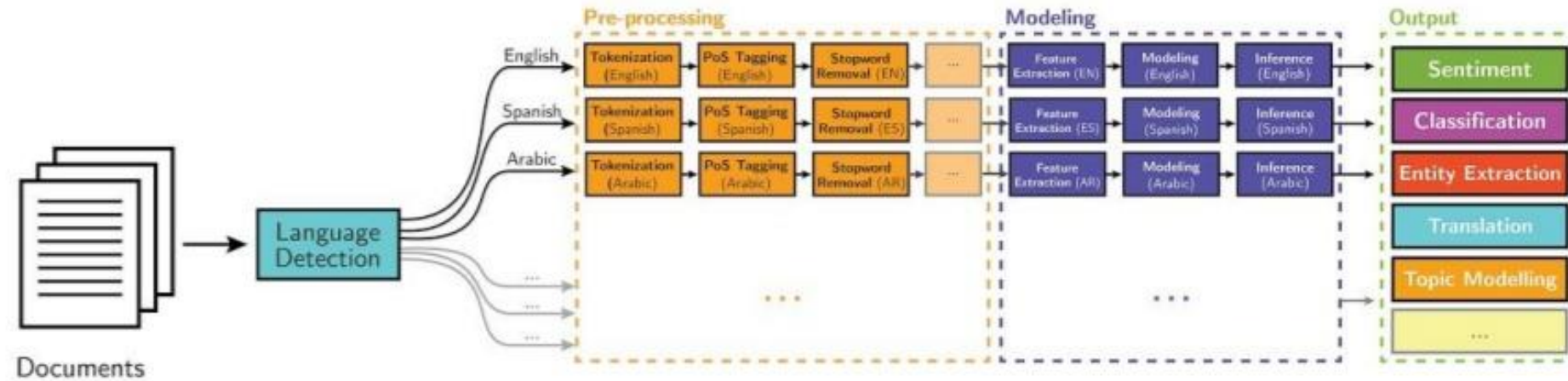
Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

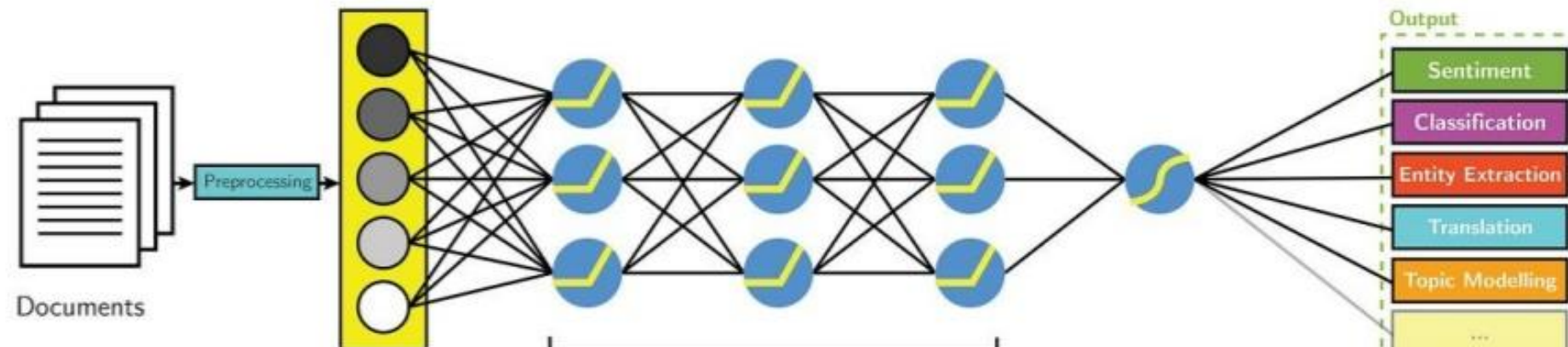
1 Introduction

Recurrent neural networks, long short-term memory [12] and gated recurrent [7] neural networks in particular, have been firmly established as state of the art approaches in sequence modeling and transduction problems such as language modeling and machine translation [29, 2, 5]. Numerous efforts have since continued to push the boundaries of recurrent language models and encoder-decoder architectures [31, 21, 13].

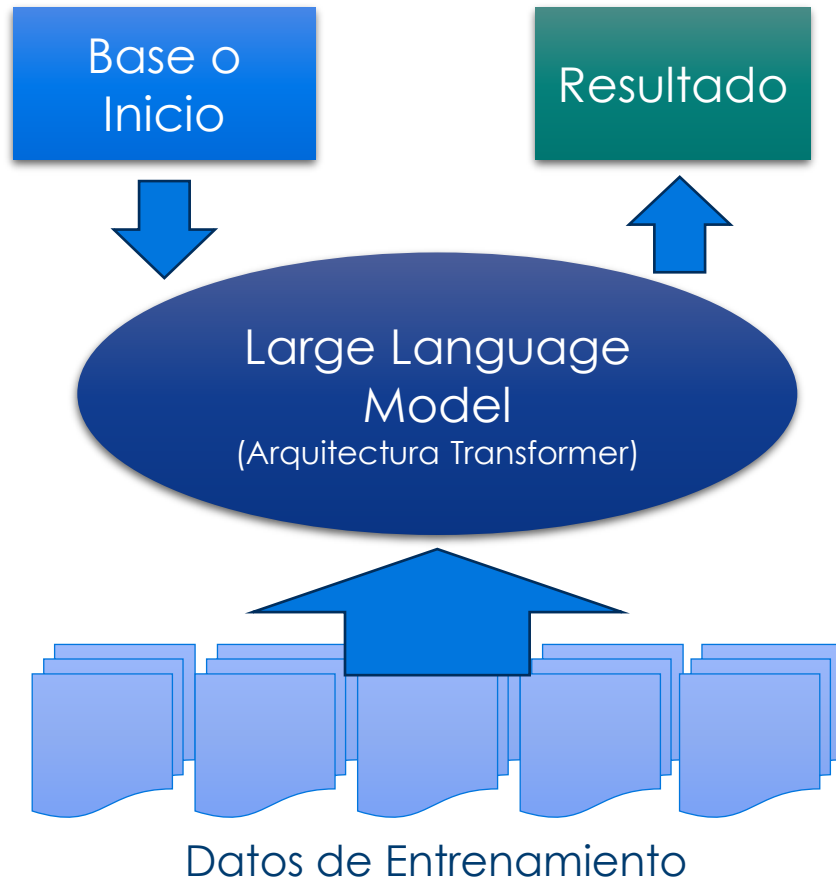
Deep learning for NLP



Deep Learning-based NLP



LLM (Large Language Models): Descripción General

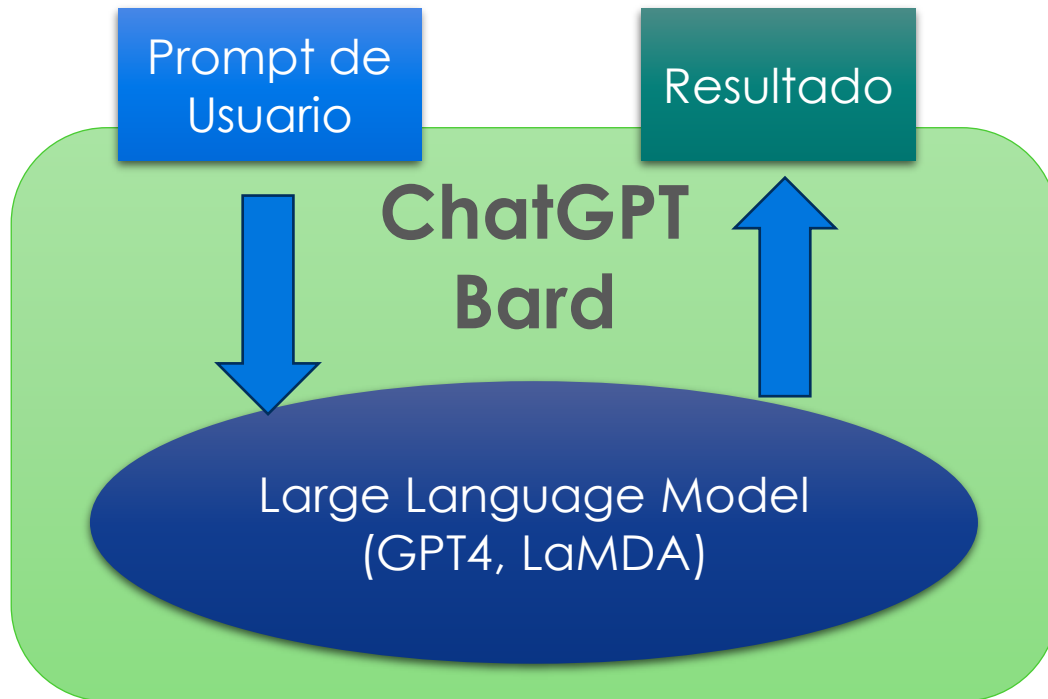


Los LLM utilizan las enormes fuentes de ejemplos de entrenamiento para construir un modelo de lenguaje capaz de entregar una secuencia de texto adecuada, según un punto de partida. Con algunas diferencias menores, hoy (2023) hay decenas de modelos entrenados con enormes cantidades de fuentes.

ALGUNOS MODELOS:

- Google: BERT (Bidirectional Encoder Representations from Transformers)
- OpenAI: GPT (2, 3, 4) (Generative Pretrained Transformer)
- Meta: LLaMA

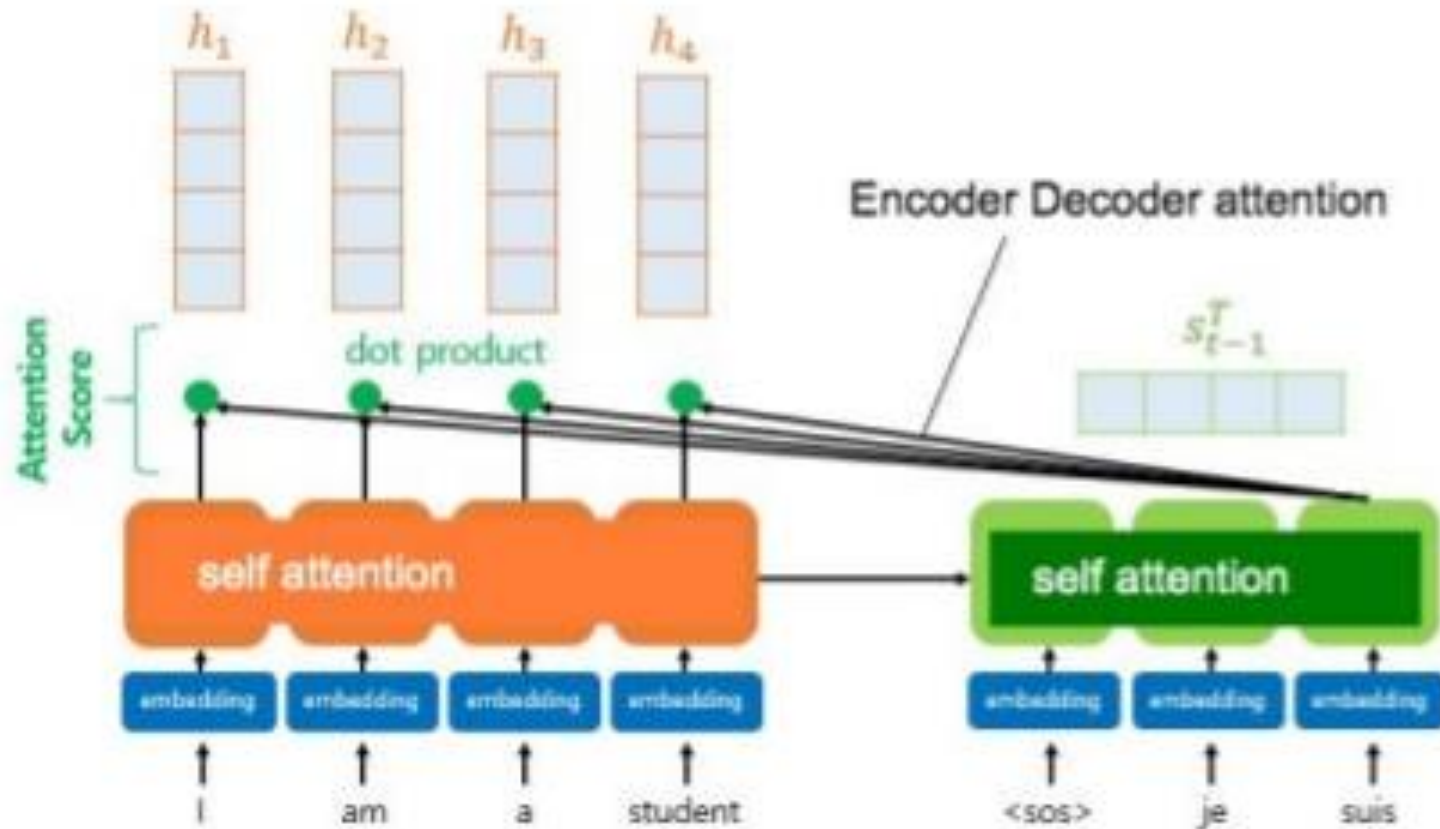
Agentes/Asistentes



Existen versiones implementadas como asistente (ChatGPT, Bard), las que reciben un prompt (o “pregunta”) de usuario, la cual es procesada para impulsar al LLM a generar una respuesta de texto que tenga sentido según el input.

El resultado siempre es una secuencia de caracteres relacionada con la entrada recibida, según la gigantesca cantidad de ejemplos que relacionan texto.

¿Por qué Transformers?



Combinar diferentes modelos de:

- Embeddings y codificación+ decodificación.
- Secuencia en la forma del mecanismo de atención.

Componentes Transformer

- Mecanismo de Atención.
 - En forma más sofisticada que las RNN, es capaz de “mantener la atención” en secuencias infinitas de elementos.
- El Transformer mismo, que incorpora:
 - Word Embeddings.
 - Encoder/Decoders.
 - Mecanismo de Atención.

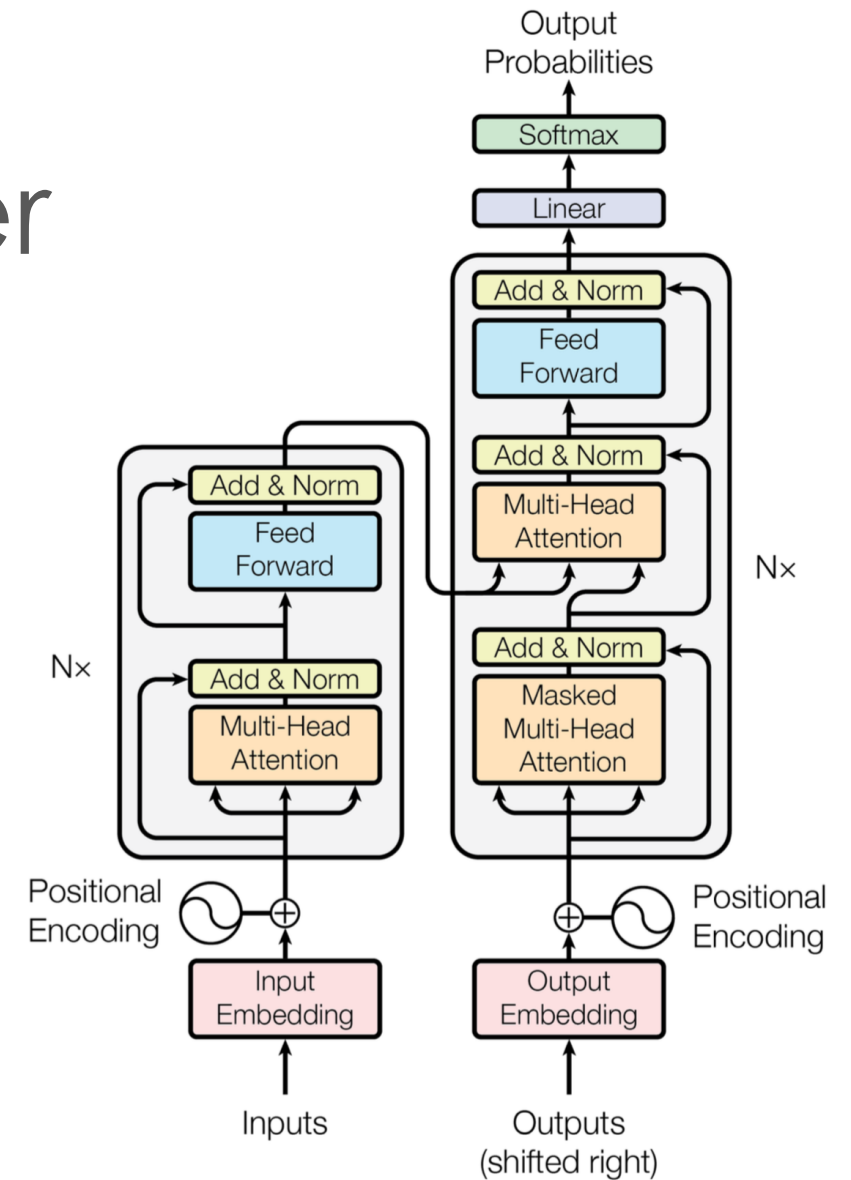
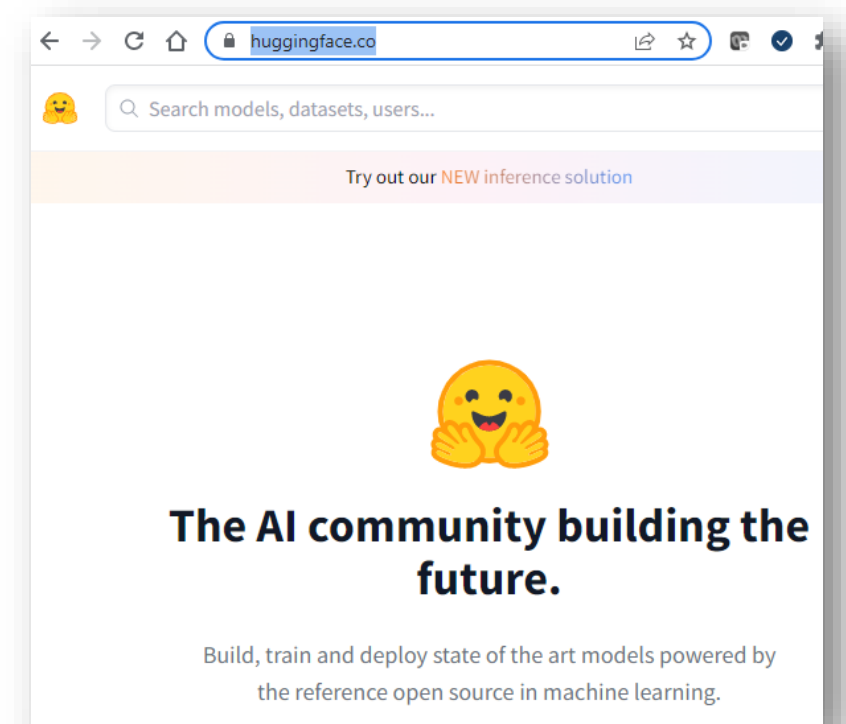


Figure 1: The Transformer - model architecture.

Hugging Face

<https://huggingface.co/>

Organización (basada en una comunidad) enfocada en el desarrollo y aplicación de Transformers. Domina arquitecturas de propósito general (BERT, GPT, RoBERTa, XLM, DistilBert, CLNet, ...) para el Entendimiento de Lenguaje Natural (NLU) y la Generación de Lenguaje Natural (NLG) sobre miles de modelos pre-entrenados en más de 100 idiomas. Hoy en día, desarrollando aplicaciones en diferentes contextos, también en ámbito de imágenes y video.





TECH \ ARTIFICIAL INTELLIGENCE \

OpenAI has published the text-generating AI it said was too dangerous to share

13

The lab says it's seen 'no strong evidence of misuse so far'

By James Vincent | Nov 7, 2019, 7:24am EST

f t SHARE

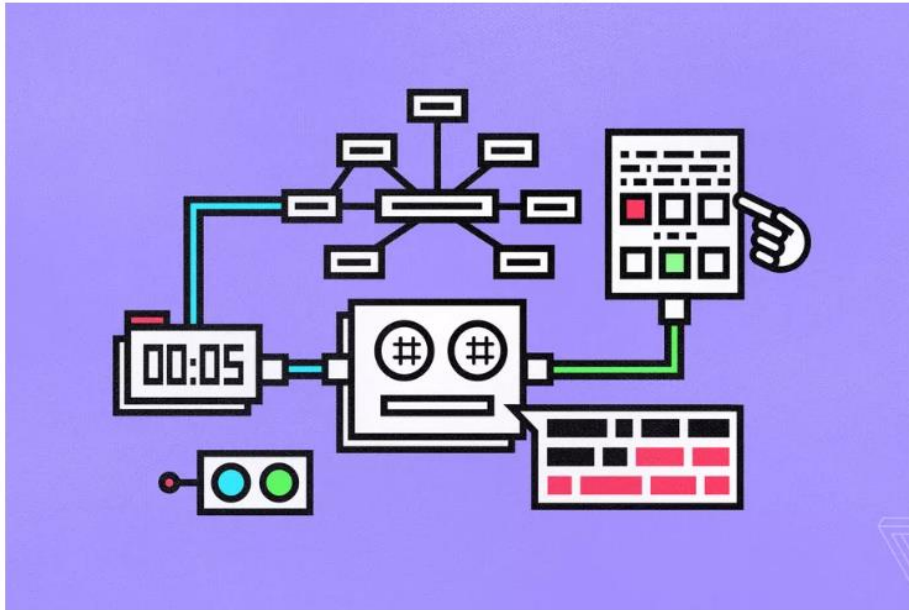


Illustration by Alex Castro / The Verge

NOVEMBER 5, 2019 • 4 MINUTE READ

GPT-2: 1.5B Release

As the final model release of GPT-2's staged release, we're releasing the largest version (1.5B parameters) of GPT-2 along with [code](#) and [model weights](#) to facilitate detection of outputs of GPT-2 models. While there have been larger language models released since August, we've continued with our original staged release plan in order to provide the community with a test case of a full staged release process. We hope that this test case will be useful to developers of future powerful models, and we're actively continuing the conversation with the AI community on responsible publication.

REPORT

GPT-2 MODEL

DETECTOR MODEL

MODEL CARD

Our findings

1. Humans find GPT-2 outputs convincing. Our partners at Cornell University surveyed people to assign GPT-2 text a credibility score across model sizes. People gave the 1.5B model a "credibility score" of 6.91 out of 10. This is marginally greater than outputs from the 774M model (6.72) and significantly above the medium 355M model (6.07). These results make us more inclined to release the 1.5B model, as the incremental increase in human-perceived credibility relative to 774M seems low.

2. GPT-2 can be fine-tuned for misuse. Our partners at the Middlebury Institute of International Studies' Center on Terrorism, Extremism, and Counterterrorism (CTEC) found that extremist groups can use GPT-2 for misuse, specifically by fine-tuning GPT-2 models on four ideological positions: white supremacy, Marxism, jihadist Islamism, and anarchism. CTEC demonstrated that it's possible to create

<https://openai.com/blog/gpt-2-1-5b-release/>

Generación de Texto

<https://talktotransformer.com/>

InferKit DEMO

10000 / 10000
weekly free characters

Sign In

El desarrollo de grandes modelos de lenguaje puede entregar
soluciones en diversos rubros como|

Generate Text

InferKit DEMO

10000 / 10000
weekly free characters

Sign In

El desarrollo de grandes modelos de lenguaje puede entregar grandes
soluciones en diversos rubros como los que tenemos en juego en el
mundo actual, o trabajamos hace algunos años en ciertos casos en la
Comunidad de Madrid y también hemos hecho algunas exportaciones
de investigaciones que están actualmente en la Universidad de Oxford. X
¿Qué pasa cuando un proyecto llega a ganar el premio grant
Guggenheim y empieza a generar fondos para sus avances, asegurando
a la Universidad de Harvard que los segu

Generate Text

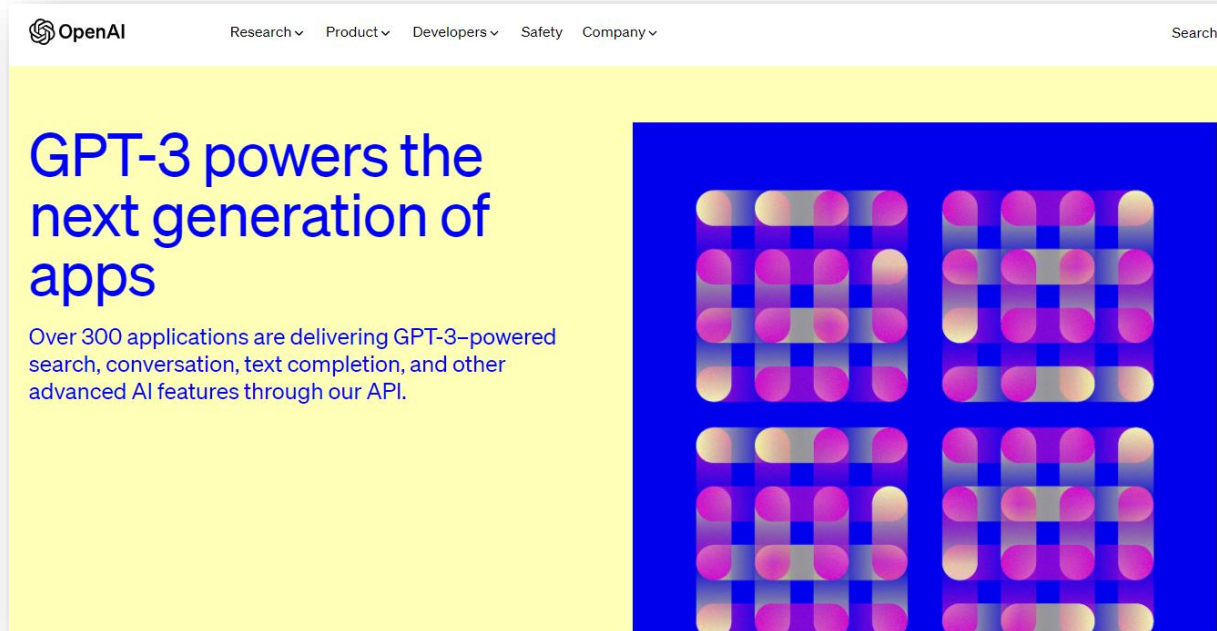




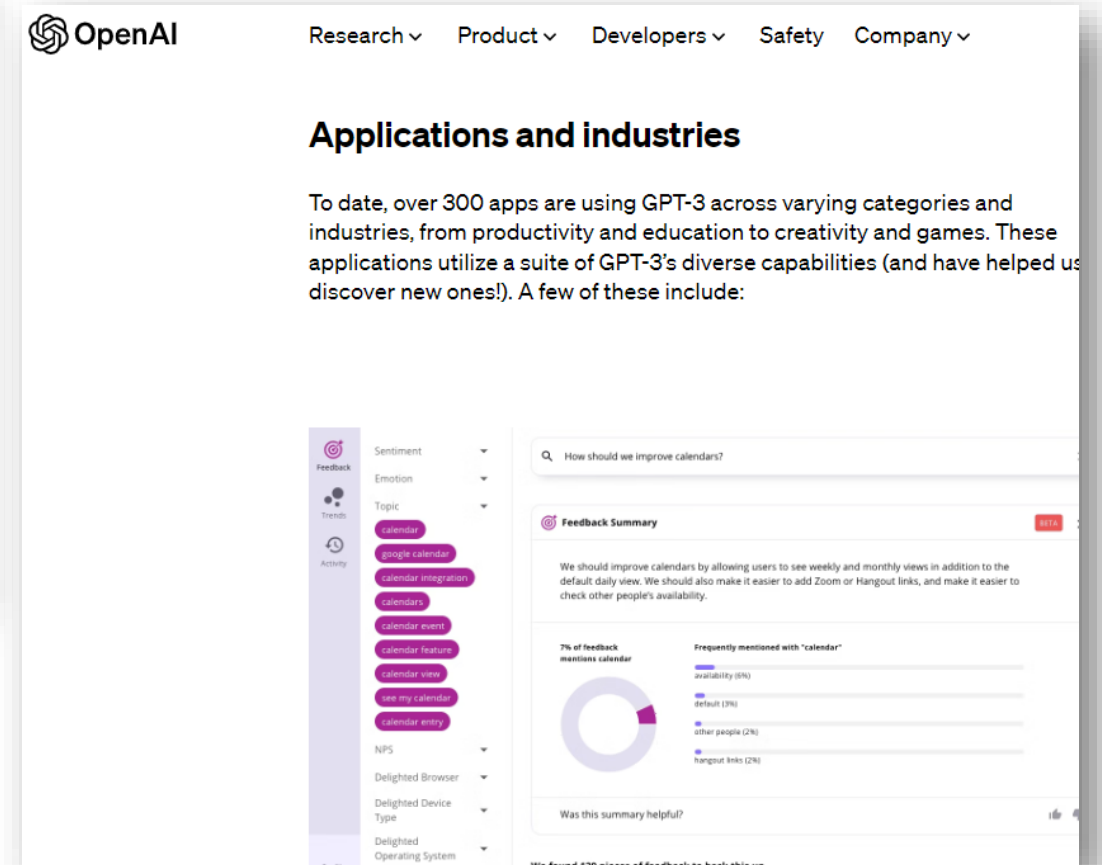
Por favor, genera argumentos legales en los que la exposición de antecedentes comprueban la disconformidad de la demanda

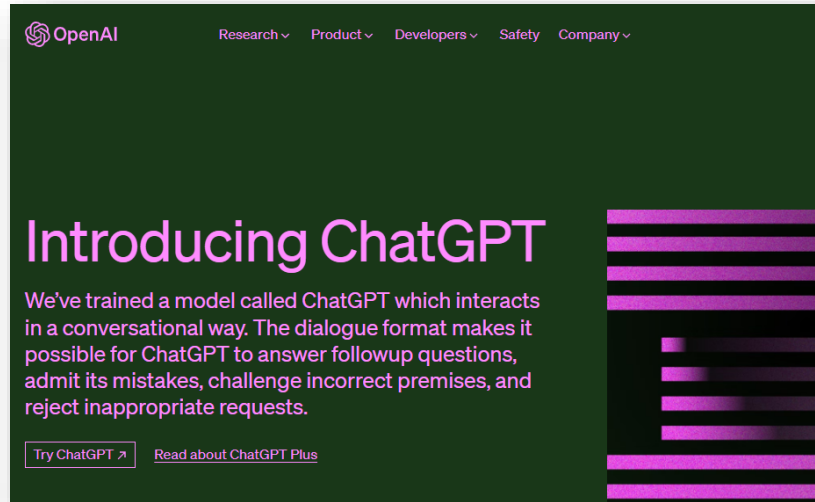


La exposición de motivos contenida en el recurso, solo da cuenta de la disconformidad que no se aplica al caso de marras el texto del artículo NUMERO N° NUMERO de la Ley sobre Impuesto a la Renta antes de su modificación por la Ley N° NUMERO -el texto posterior a la modificación trata sólo el reajuste de los derechos en sociedades de personas-, ello importó que para los jueces del grado esta materia quedara entonces gobernada únicamente por el artículo NUMERO N° NUMERO.
En efecto, el arbitrio se extiende sobre...



<https://openai.com/blog/gpt-3-apps>





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



Here are a few suggestions to improve the code:



1. The variable ``randomNum`` is being instantiated with the current second of the date time, this is not a good seed value as it will always be the same when the function is called within the same second. Instead, you can use the ``Guid.NewGuid().GetHashCode()`` as a seed value for the random generator.
2. The variable ``questionsPerGroup`` is being used to store the quantity of questions per group, it could be replaced with a ``Dictionary<int,int>`` where the key is the group number and the value is the quantity of questions.
3. The variable ``sortedQuestions`` is being used to store a random question from each group, it could be replaced with a ``Dictionary<int,TestQuestion>`` where the key is the group number and the value is the selected question.
4. The variable ``questionIndexOffset`` is being used to track the index offset of each group, it could be replaced with a variable ``currentIndex`` that keeps track of the current index when iterating through the list of questions.
5. Instead of removing the item from the list ``sortedQuestions`` when shuffling, it's better

to use the ``List.Shuffle()`` method to even the current item with a randomly selected one

Methods

We trained this model using Reinforcement Learning from Human Feedback (RLHF), using the same methods as InstructGPT, but with slight differences in the data collection setup. We trained an initial model using supervised fine-tuning: human AI trainers provided conversations in which they played both sides—the user and an AI assistant. We gave the trainers access to model-written suggestions to help them compose their responses. We mixed this new dialogue dataset with the InstructGPT dataset, which we transformed into a dialogue format.

To create a reward model for reinforcement learning, we needed to collect comparison data, which consisted of two or more model responses ranked by quality. To collect this data, we took conversations that AI trainers had with the chatbot. We randomly selected a model-written message, sampled several alternative completions, and had AI trainers rank them. Using these reward models, we can fine-tune the model using Proximal Policy Optimization. We performed several iterations of this process.

Step 1

Collect demonstration data and train a supervised policy.

A prompt is sampled from our prompt dataset.

Explain reinforcement learning to a 6 year old.

A labeler demonstrates the desired output behavior.

We give treats and punishments to teach...

This data is used to fine-tune GPT-3.5 with supervised

SFT

Step 2

Collect comparison data and train a reward model.

A prompt and several model outputs are sampled.

Explain reinforcement learning to a 6 year old.

A. Reinforcement learning is the...

B. Define reward...

C. Explain to me...

D. No give them and punishments to teach...

A labeler ranks the outputs from best to worst.

U > C > A > B

Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

A new prompt is sampled from the dataset.

Write a story about otters.

The PPO model is initialized from the supervised policy.

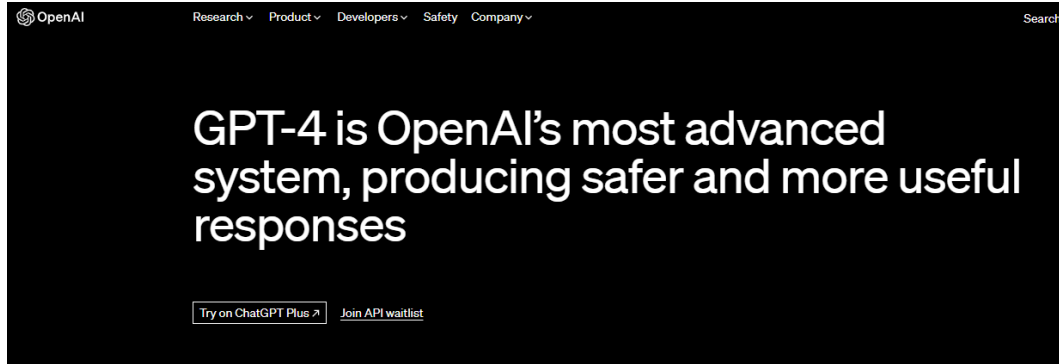
PPO

The policy generates an output.

Once upon a time...

The reward model calculates a reward

RM



<https://openai.com/product/gpt-4>

OpenAI Research Product Developers Safety Company Search

GPT-4 surpasses ChatGPT in its advanced reasoning capabilities.

ChatGPT

Input
Andrew is free from 11 am to 3 pm, Joanne is free from noon to 2 pm and then 3:30 pm to 5 pm. Hannah is available at noon for half an hour, and then 4 pm to 6 pm. What are some options for start times for a 30 minute meeting for Andrew, Hannah, and Joanne?

Output
Here is a list of their availability:
- Andrew: 11 am to 3 pm
- Joanne: noon to 2 pm, and 3:30 pm to 5 pm
- Hannah: noon to 12:30 pm, and 4 pm to 6 pm

Based on their availability, there is a 30-minute window where all three of them are available, which is from 4 pm to 4:30 pm. So, the meeting can be scheduled at 4 pm.

GPT-4

Input
Andrew is free from 11 am to 3 pm, Joanne is free from noon to 2 pm and then 3:30 pm to 5 pm. Hannah is available at noon for half an hour, and then 4 pm to 6 pm. What are some options for start times for a 30 minute meeting for Andrew, Hannah, and Joanne?

Output
Andrew: 11 am - 3 pm
Joanne: 12 pm - 2 pm, 3:30 pm - 5 pm
Hannah: 12 pm - 12:30 pm, 4 pm - 6 pm

Common availability for a 30-minute meeting: 12 pm - 12:30 pm

GPT-4 outperforms ChatGPT by scoring in higher approximate percentiles among test-takers.

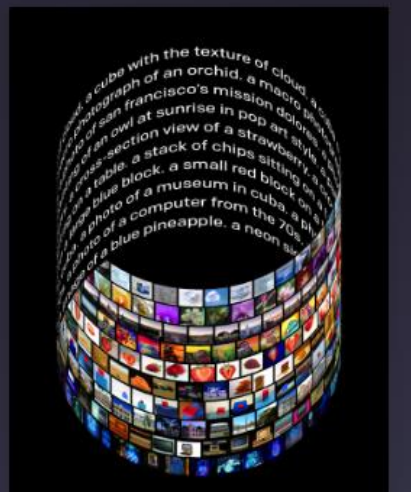
Uniform Bar Exam		Biology Olympiad	
10th	90th	31st	99th
ChatGPT	GPT-4	ChatGPT	GPT-4 (with vision)

DALL-E: una aplicación de GPT-3

DALL-E: Creating Images from Text

We've trained a neural network called DALL-E that creates images from text captions for a wide range of concepts expressible in natural language.

January 5, 2021



DALL-E es una versión de GPT-3 de 12 mil millones de parámetros, entrenada para generar imágenes desde descripciones de texto, usando pares texto-imagen para aprender. Es capaz de combinar conceptos de ejemplos separados, produciendo resultados sorprendentes sobre tópicos disjuntos, generando texto y aplicando transformaciones a imágenes existentes (conocidas en entrenamiento).

arXiv:2102.12092v2 [cs.CV] 26 Feb 2021

Zero-Shot Text-to-Image Generation

Aditya Ramesh¹ Mikhail Pavlov¹ Gabriel Goh¹ Scott Gray¹
Chelsea Voss¹ Alec Radford¹ Mark Chen¹ Ilya Sutskever¹

Abstract

Text-to-image generation has traditionally focused on finding better modeling assumptions for training on a fixed dataset. These assumptions might involve complex architectures, auxiliary losses, or side information such as object part labels or segmentation masks supplied during training. We describe a simple approach for this task based on a transformer that autoregressively models the text and image tokens as a single stream of data. With sufficient data and scale, our approach is competitive with previous domain-specific models when evaluated in a zero-shot fashion.

1. Introduction

Modern machine learning approaches to text to image synthesis started with the work of Mansimov et al. (2015), who showed that the DRAW Gregor et al. (2015) generative model, when extended to condition on image captions, could also generate novel visual scenes. Reed et al. (2016b) later demonstrated that using a generative adversarial network (Goodfellow et al., 2014), rather than a recurrent variational auto-encoder, improved image fidelity. Reed et al. (2016b) showed that this system could not only generate objects with recognizable properties, but also could *zero-shot* generalize to held-out categories.

Over the next few years, progress continued using a combination of methods. These include improving the generative model architecture with modifications like multi-scale generators (Zhang et al., 2017; 2018), integrating attention and auxiliary losses (Xu et al., 2018), and leveraging additional sources of conditioning information beyond just text (Reed et al., 2016a; Li et al., 2019; Koh et al., 2021).

Separately, Nguyen et al. (2017) propose an energy-based framework for conditional image generation that obtained a large improvement in sample quality relative to contemporary methods. Their approach can incorporate pretrained discriminative models, and they show that it is capable of performing text-to-image generation when applied to a cap-

¹OpenAI, San Francisco, California, United States. Correspondence to {aditya, mikhail, gabriel, scott, chelsea, alec, mark, ilya}@openai.com



Figure 1. Comparison of original images (top) and reconstructions from the discrete VAE (bottom). The encoder downsamples the spatial resolution by a factor of 8. While details (e.g., the texture of the cat's fur, the writing on the storefront, and the thin lines in the illustration) are sometimes lost or distorted, the main features of the image are still typically recognizable. We use a large vocabulary size of 8192 to mitigate the loss of information.

tioning model pretrained on MS-COCO. More recently, Cho et al. (2020) also propose a method that involves optimizing the input to a pretrained cross-modal masked language model. While significant increases in visual fidelity have occurred as a result of the work since Mansimov et al. (2015), samples can still suffer from severe artifacts such as object distortion, illogical object placement, or unnatural blending of foreground and background elements.

Recent advances fueled by large-scale generative models suggest a possible route for further improvements. Specifically, when compute, model size, and data are scaled carefully, autoregressive transformers (Vaswani et al., 2017) have achieved impressive results in several domains such as text (Radford et al., 2019), images (Chen et al., 2020), and audio (Dhariwal et al., 2020).

By comparison, text-to-image generation has typically been evaluated on relatively small datasets such as MS-COCO and CUB-200 (Welinder et al., 2010). Could dataset size and model size be the limiting factor of current approaches? In this work, we demonstrate that training a 12-billion param-



Ejemplos DALL-E

TEXT PROMPT an armchair in the shape of an avocado. . . .

AI-GENERATED
IMAGES



Edit prompt or view more images+

TEXT PROMPT a store front that has the word 'openai' written on it. . . .

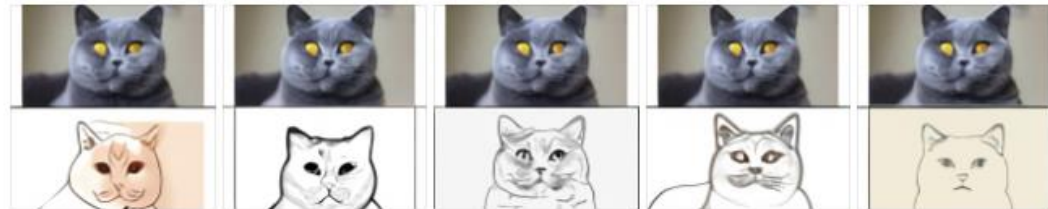
AI-GENERATED
IMAGES



Edit prompt or view more images+

TEXT & IMAGE
PROMPT the exact same cat on the top as a sketch on the bottom

AI-GENERATED
IMAGES



openai.com/blog/dall-e/

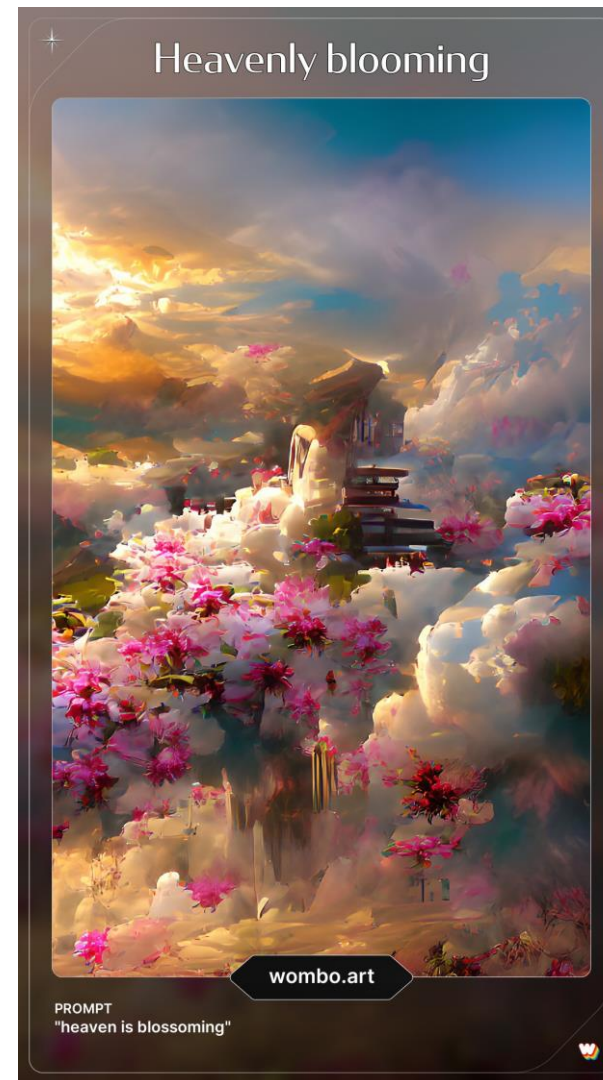


wombo.art



wombo.art

PROMPT
"mountains in winter"



wombo.art

PROMPT
"heaven is blossoming"

¿Es ChatGPT la IA definitiva?

ChatGPT is not all you need. A State of the Art Review of large Generative AI models

Roberto Gozalo-Brizuela, Eduardo C. Garrido-Merchán

Quantitative Methods Department, Universidad Pontificia Comillas, Madrid, Spain
201905616@alu.comillas.edu, ecgarrido@icade.comillas.edu

Abstract. During the last two years there has been a plethora of large generative models such as ChatGPT or Stable Diffusion that have been published. Concretely, these models are able to perform tasks such as being a general question and answering system or automatically creating artistic images that are revolutionizing several sectors. Consequently, the implications that these generative models have in the industry and society are enormous, as several job positions may be transformed. For example, Generative AI is capable of transforming effectively and creatively texts to images, like the DALL-E-2 model; text to 3D images, like the Dreamfusion model; images to text, like the Flamingo model; texts to video, like the Phenaki model; texts to audio, like the AudioLM model; texts to other texts, like ChatGPT; texts to code, like the Codex model; texts to scientific texts, like the Galactica model or even create algorithms like AlphaTensor. This work consists on an attempt to describe in a concise way the main models and sectors that are affected by generative AI and to provide a taxonomy of the main generative models published recently.

1 Introduction

Generative AI refers to artificial intelligence that can generate novel content, rather than simply analyzing or acting on existing data like expert systems [23]. In particular, expert systems contained knowledge bases and an inference engine

¿Reemplaza a Google con mejores resultados?

NO. ChatGPT no se actualiza con nuevas referencias en la web en cada instante. Está entrenado con un cuerpo de conocimiento enorme, pero acotado a un instante de tiempo.

¿Va a reemplazar empleos tradicionales?

NO. ChatGPT no es capaz de reemplazar la capacidad lógica de los profesionales en sus especialidades, sin embargo es capaz de simplificarles el trabajo, avanzando en tareas más simples, como escribir texto para un informe, una noticia, un argumento legal, un análisis financiero.

¿Es ChatGPT la IA Definitiva?



Yann LeCun
@ylecun

...

To be clear: I'm not criticizing OpenAI's work nor their claims.

I'm trying to correct a *perception* by the public & the media who see chatGPT as this incredibly new, innovative, & unique technological breakthrough that is far ahead of everyone else.

It's just not.

[Traducir Tweet](#)

Personas relevantes



Yann LeCun
@ylecun

Siguiendo

Professor at NYU. Chief AI Scientist at Meta. Researcher in AI, Machine Learning, Robotics, etc. ACM Turing Award Laureate.



Gracias



rsandova@ing.puc.cl
rodrigo@RSolver.com



@RSandovalSolver



/in/RodrigoSandoval

www.RodrigoSandoval.net
www.RSolver.com