Microcredencial Universitaria

Financiado por la Unión Europea NextGenerationEU

GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

Plan de Recuperación, Transformación y Resiliencia

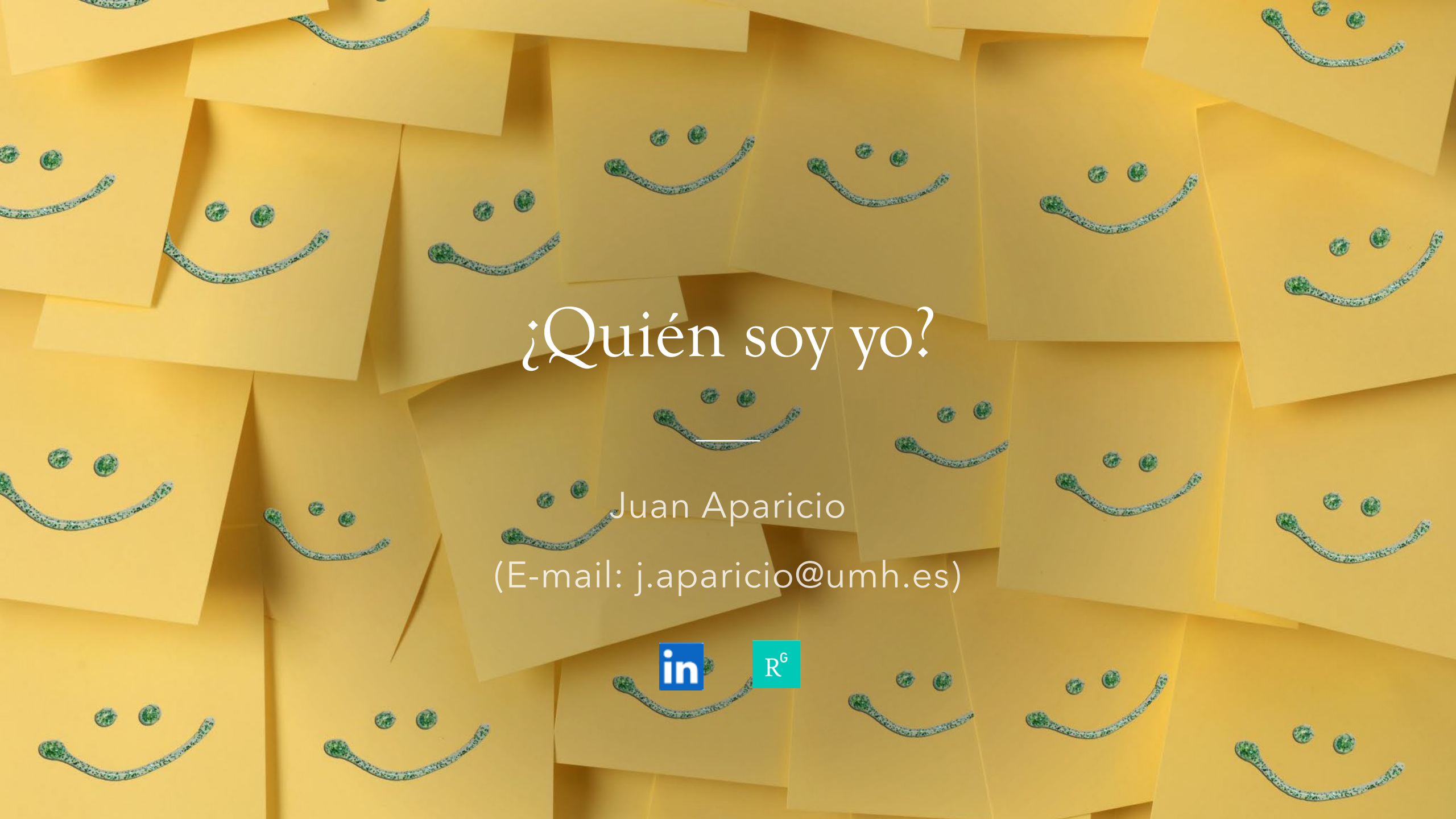
GENERALITAT VALENCIANA Conselleria de Educación, Cultura y Universidades

GVANEXT Fondos Next Generation en la Comunitat Valenciana



Tema 1. Introducción a la IA Generativa

Juan Aparicio & Xavier Barber (UMH)



¿Quién soy yo?

Juan Aparicio

(E-mail: j.aparicio@umh.es)

Juan Aparicio



- Catedrático del área de Estadística e IO de la UMH (y estadístico de formación)
- Director del Centro de Investigación Operativa
- Miembro de ValgrAI (Valencian Graduate School and Research Network of Artificial Intelligence)
- En investigación: aplicación de técnicas de machine learning para la mejora de la eficiencia y la productividad
- Transferencia de conocimiento: Mutualidad de Levante, Ayto. de Elche, Consellerías, MILMOH, Heort Salud (Colombia), I/O Research (USA), …
- Consultor de IA para Caja Ingenieros, constructora LarCosta, Palomares Consultores, …
- Colaborador del Ministerio del Interior (Secretaría de Estado de Seguridad), la Policía Nacional y la Guardia Civil con el proyecto VioGen.
- Cursos de formación de IA impartidos: Policías locales de la CV, AVECAL, Ayuntamiento de Dénia, Ayuntamiento de Xàbia, CREAMA, …
- Co-organizador del Palmeratech 2024 y 2025 con más de 400 participantes.
- Participación en mesas redondas y entrevistas sobre IA Generativa (El País) y co-organizador del PalmeraTech

Podcast en Youtube

@automatiza2.podcast





1. Introducción a la IA Generativa

2. Cómo funciona un LLM


State of the Art de los LLMs

Modelos de IA Generativa y sus empresas

deepseek

Apple
CLaRa
7B

Apple Intelligence

NVIDIA
NEMOTRON





OpenAI
ChatGPT

Fara-7B
Computer Use

Microsoft
Copilot







Microsoft
Phi-4

MISTRAL
AI_

Gemini



Moonshot AI
Kimi K2.5
AGENT SWARM

Amazon Nova

perplexity

Gemma



Algunos ejemplos que te explotan la cabeza

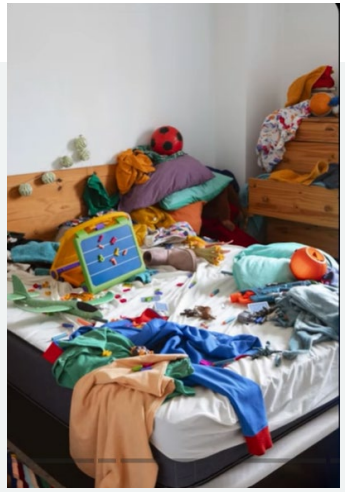
Generación de imágenes y edición

Y por supuesto……



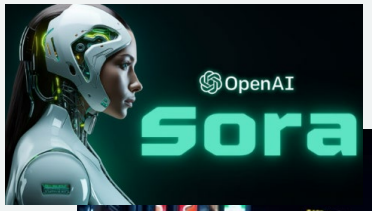


Nano Banana 2: habilidades profesionales, mucho más rápido





Ejemplos de creación de vídeo





OpenAI
Sora 2

PIKA LABS











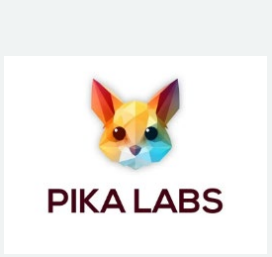




Novedades – Generación de Video

















Open Source



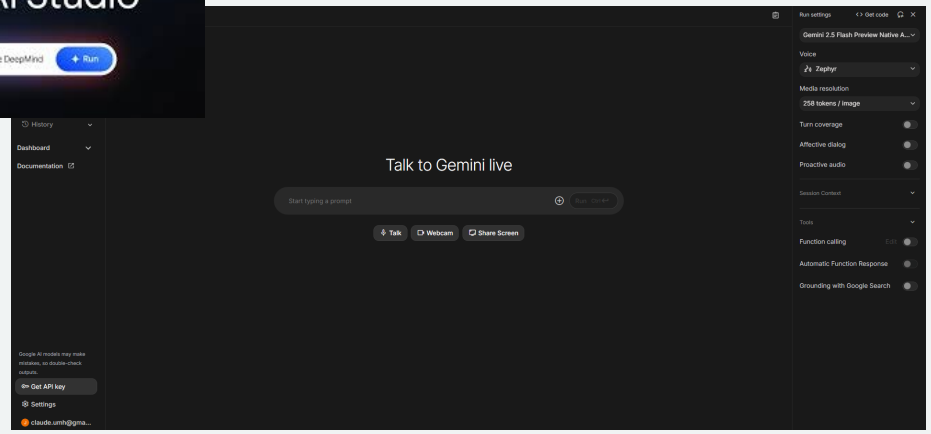


Agentes avanzados con voz

















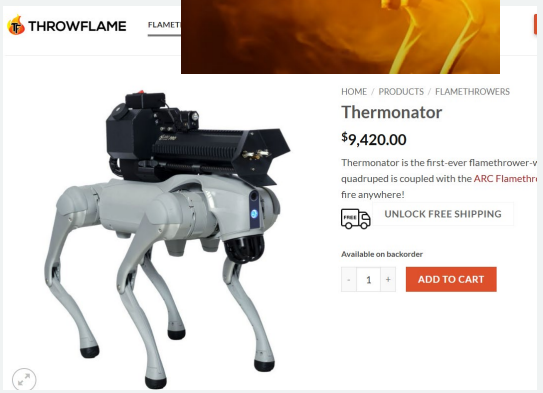




Robots






THROWFLAME
HOME / PRODUCTS / FLAMETHROWERS
Thermonator
$9,420.00
UNLOCK FREE SHIPPING
Available on backorder
ADD TO CART

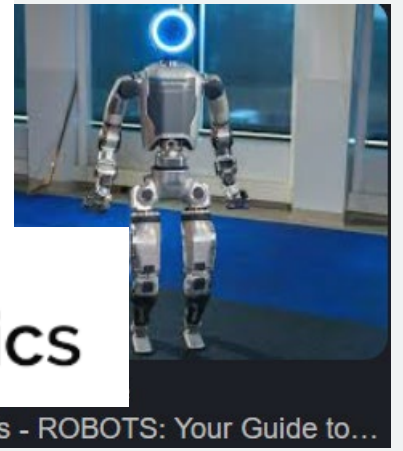

Parametric Architecture
NEO by 1X, The World's First Consum…


Pasión por el despacho - Centro de Innovación …
IRON, el robot humanoide de XPENG que …


Ineltec
Unitree G1: El nuevo competidor…

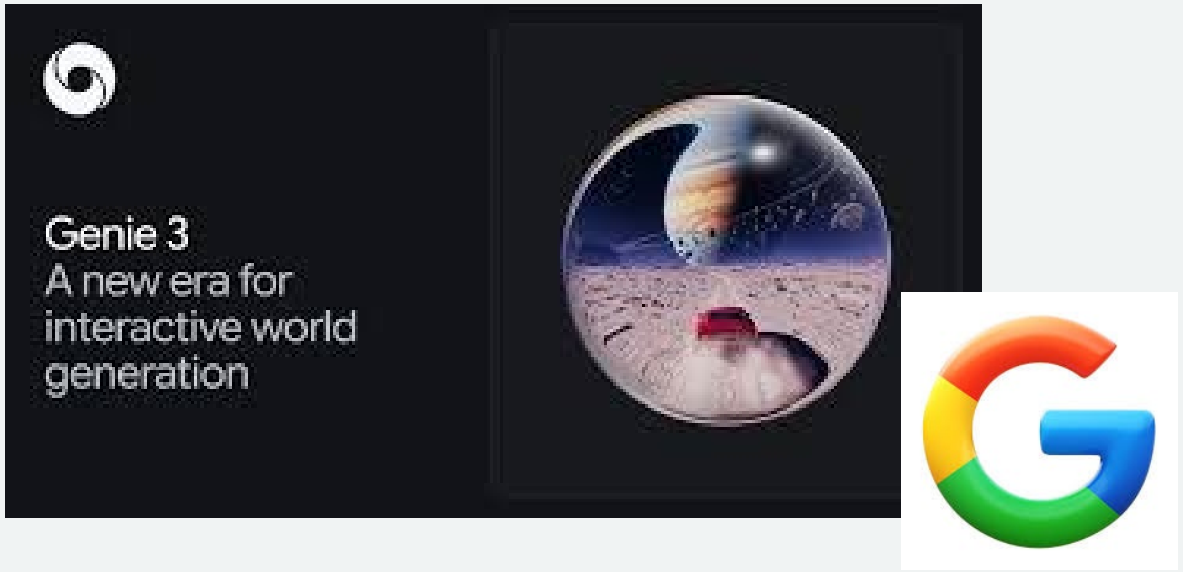

Optimus - Gen 2
Shop4Tesla
Optimus robot: Tesla se basa en el entrenamie…


Boston Dynamics


Atlas - ROBOTS: Your Guide to…







El futuro:
¿Robots + LLMs?

¿Genie 3 y simuladores de mundos de NVIDIA para entrenar a robots?

2025: El año de los Agentes de IA (pero llegó en 2026)



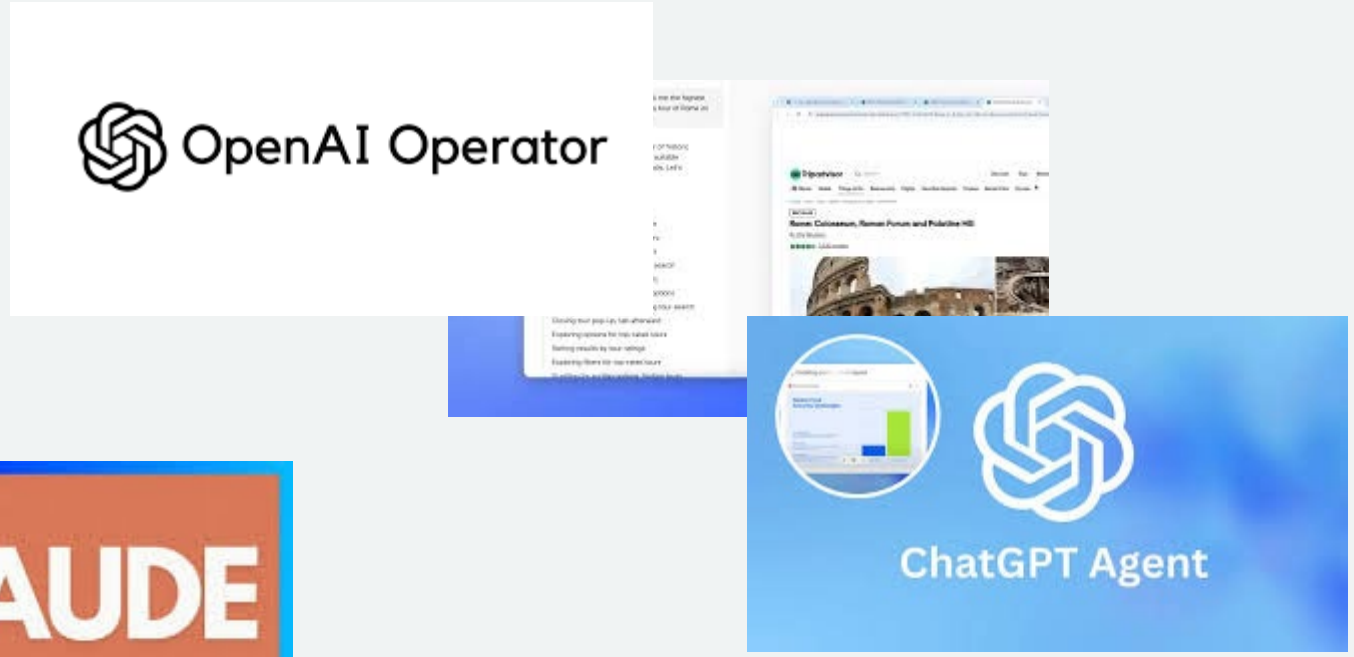
Un **agente de inteligencia artificial** es un sistema que **percibe su entorno**, **toma decisiones** y **actúa de forma autónoma** para cumplir un objetivo específico.

Características clave:

🔍 **Observa**: recibe información del entorno.
🧠 **Razona**: analiza la información y decide qué hacer.
⚙️ **Actúa**: ejecuta tareas automáticamente.
🔄 **Aprende**: puede mejorar su comportamiento con el tiempo.

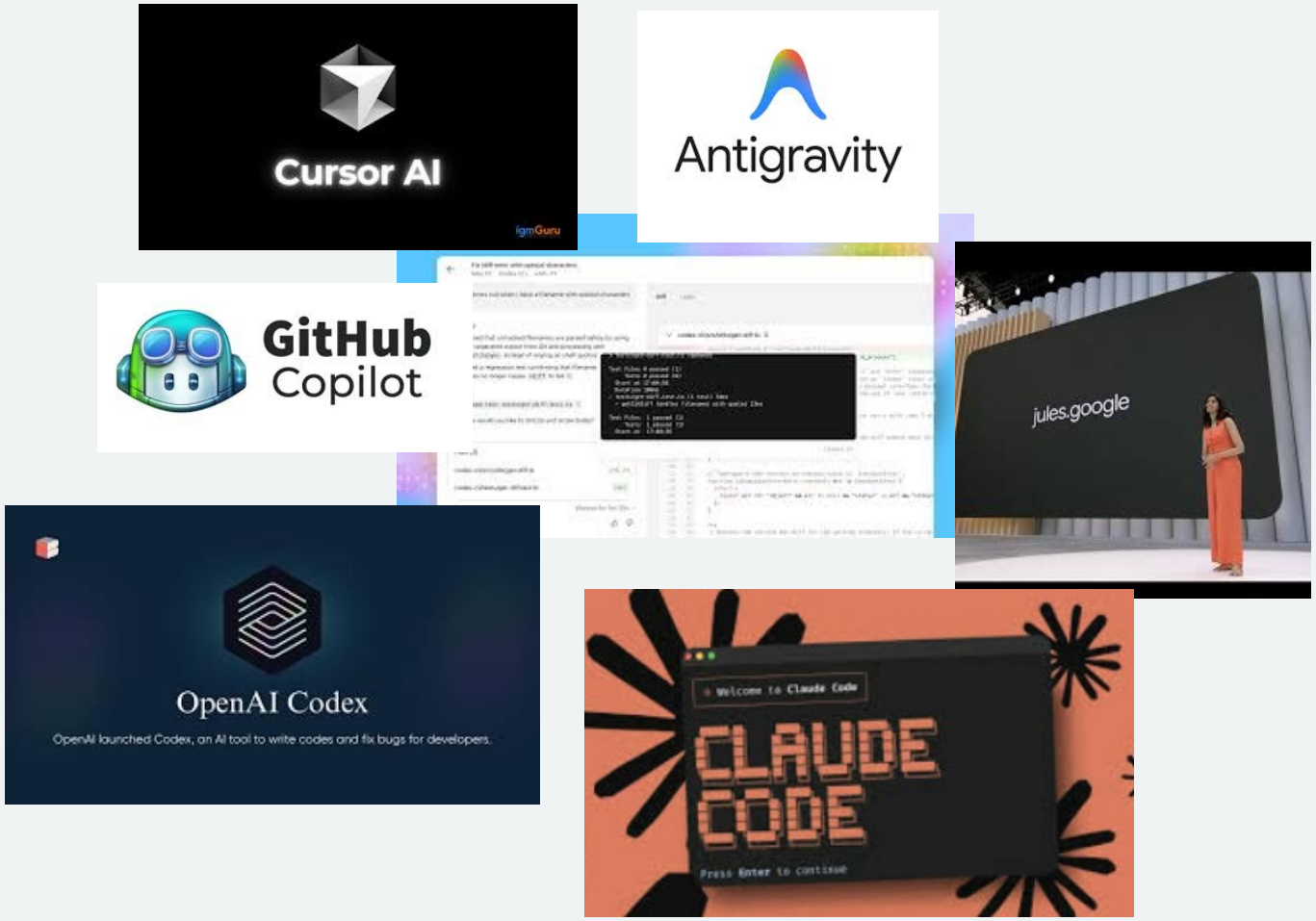
Agentes de IA – Inicios (2024-2025)

manus
The general AI agent





Agentes de IA – Para programar



Agentes de IA – Navegadores

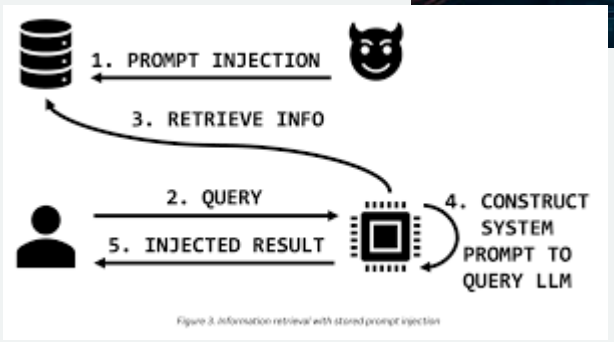












Agentes de IA – Ya casi todos tienen agentes en sus plataformas que actúan en la Web

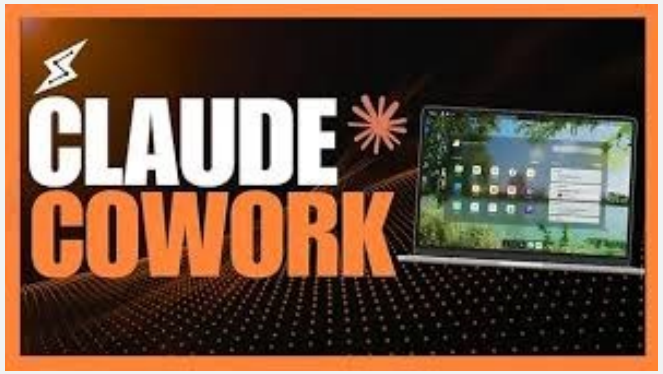








Agentes de IA – sorpresas recientes





Peter Steinberger







Peter Steinberger acaba de fichar por OpenAI. Sam Altman lo ha anunciado en X: "Peter Steinberger se une a OpenAI para impulsar la próxima generación de agentes personales."

Wearables



Meta





Grandes centros de computación, Chips y entrenamiento en el espacio

- Chips:
  - NVIDIA
  - TPU de Google
  - LPU de Groq
  - WSE de Cerebras
  - Microsoft



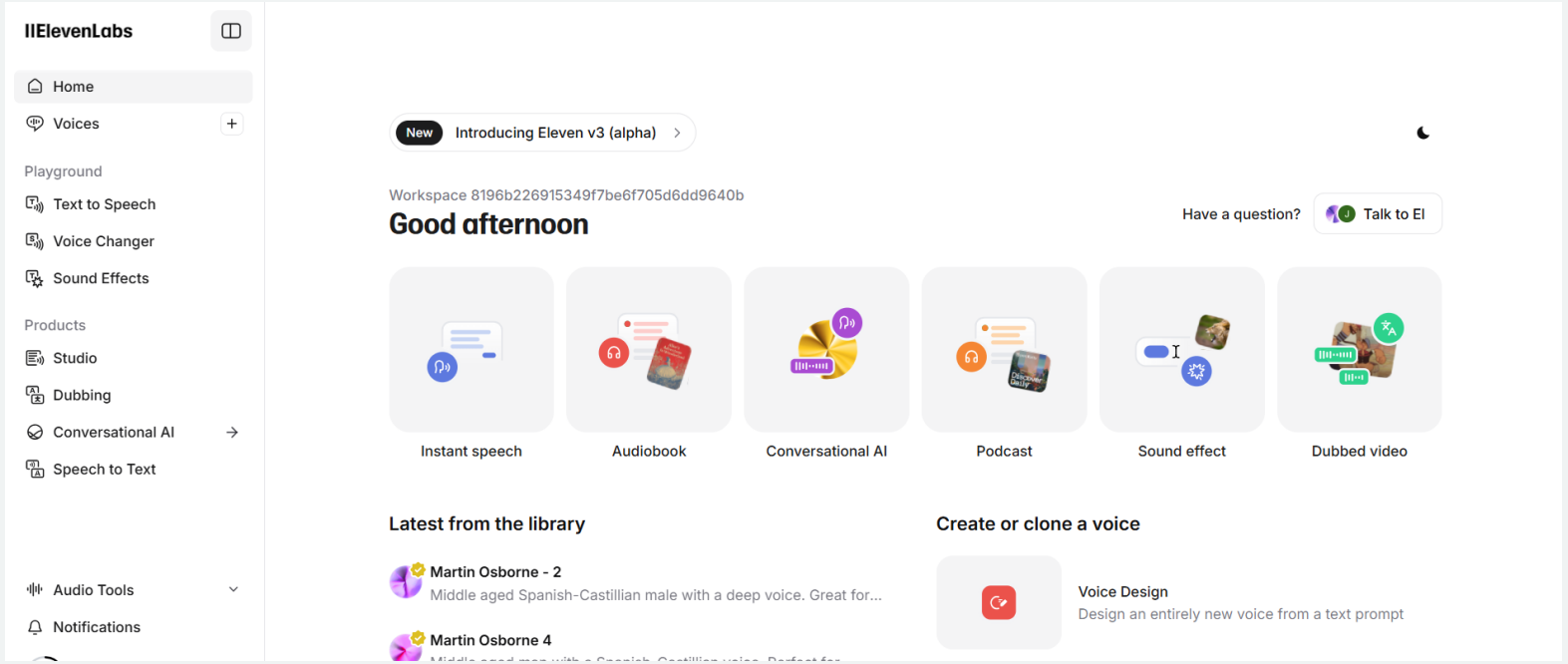


🌞 ENERGÍA → El cómputo en órbita puede tirar de energía solar casi ilimitada en vez de seguir chupando red eléctrica y agua en la Tierra

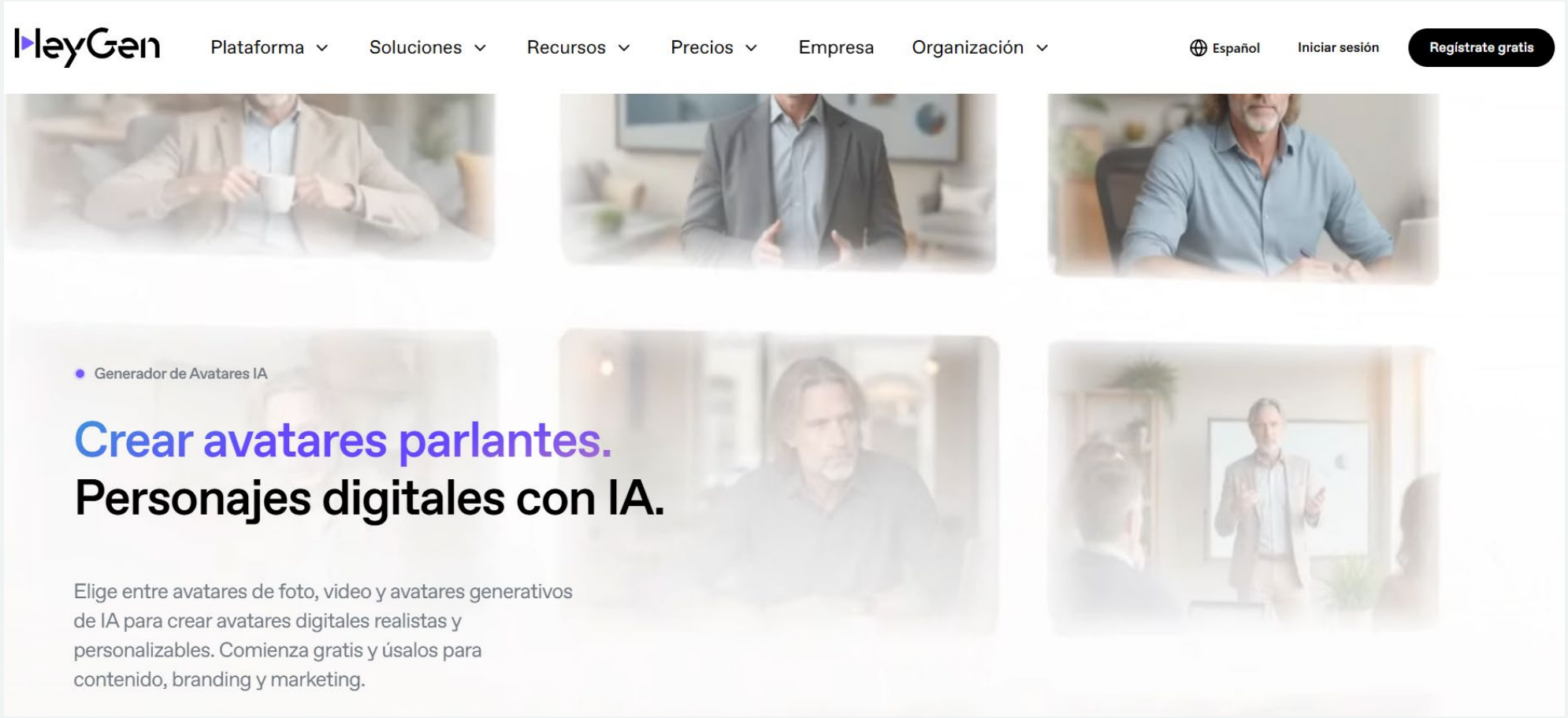
🥶 FRÍO → Los data centers en el espacio evitan muchos de los problemas de refrigeración y de uso de suelo que tienen los centros de datos en tierra

Clonado de voz: ElevenLabs

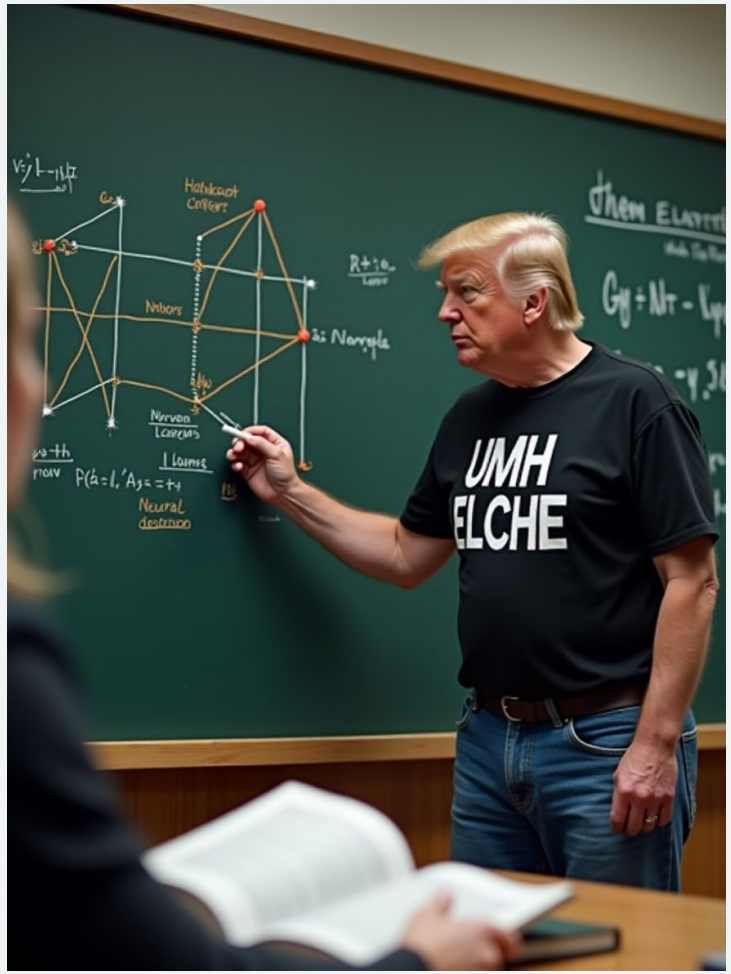
IIElevenLabs

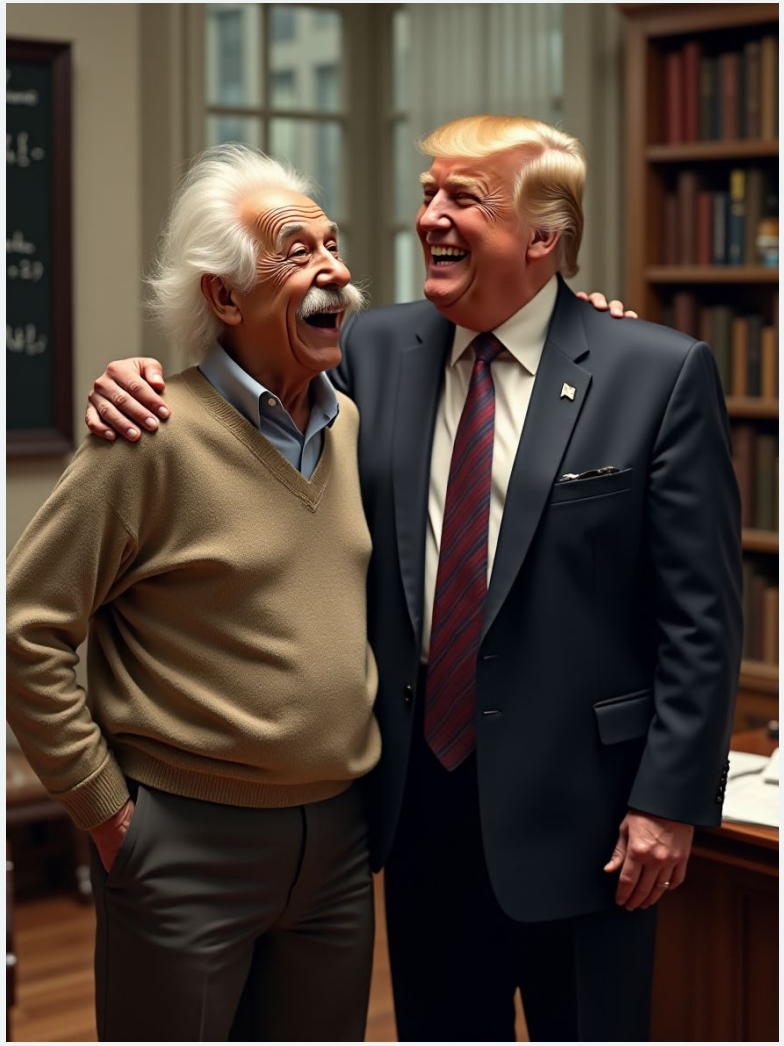


Avatares


HeyGen
Plataforma
Soluciones
Recursos
Precios
Empresa
Organización
Español
Iniciar sesión
Regístrate gratis
Generador de Avatares IA
Crear avatares parlantes.
Personajes digitales con IA.
Elige entre avatares de foto, video y avatares generativos de IA para crear avatares digitales realistas y personalizables. Comienza gratis y úsalos para contenido, branding y marketing.

Relacionado con eso...Deepfakes







Suplantan con IA la voz del CEO de Ferrari





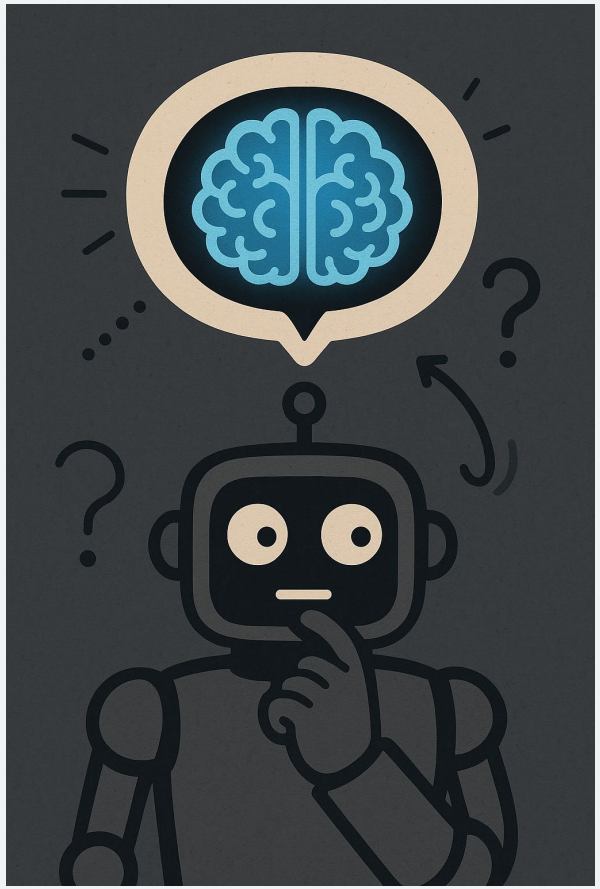
Sam Almant tiene un plan



Chatbots con contextos gigantes



Modelos razonadores



Prompt:

En una mesa hay tres cajas: una roja, una azul y una verde. Una de ellas contiene un tesoro, y las otras dos están vacías. Cada caja tiene una etiqueta, pero solo una de las etiquetas es verdadera y las otras dos son falsas.

Las etiquetas dicen lo siguiente:

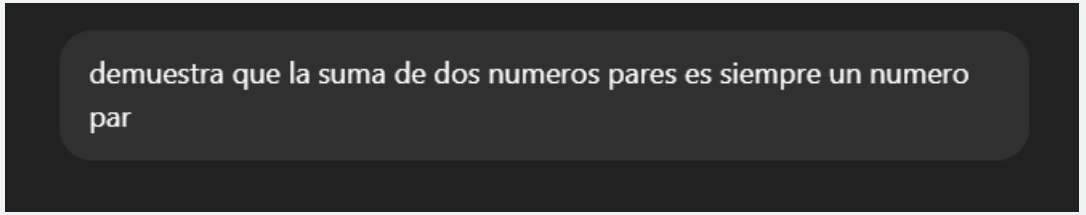
Caja Roja: "El tesoro no está en la caja azul."
Caja Azul: "El tesoro no está aquí."
Caja Verde: "El tesoro está en la caja azul."
¿En qué caja está el tesoro?

Solución: Caja Azul



No basta con hablar del *tamaño del modelo* (número de parámetros) o de los *tokens de entrada/salida*. También importa cuánto cómputo se le permite usar mientras "piensa" durante la inferencia.

Otro prompt:





Cómo se construye un TRANSFORMER

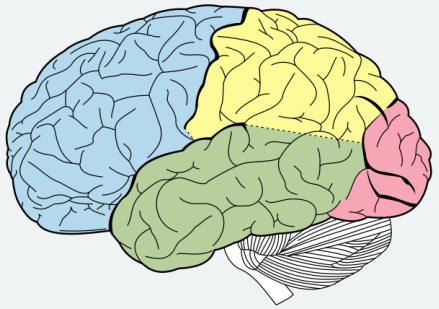
¿Cómo se construye un modelo LLM?

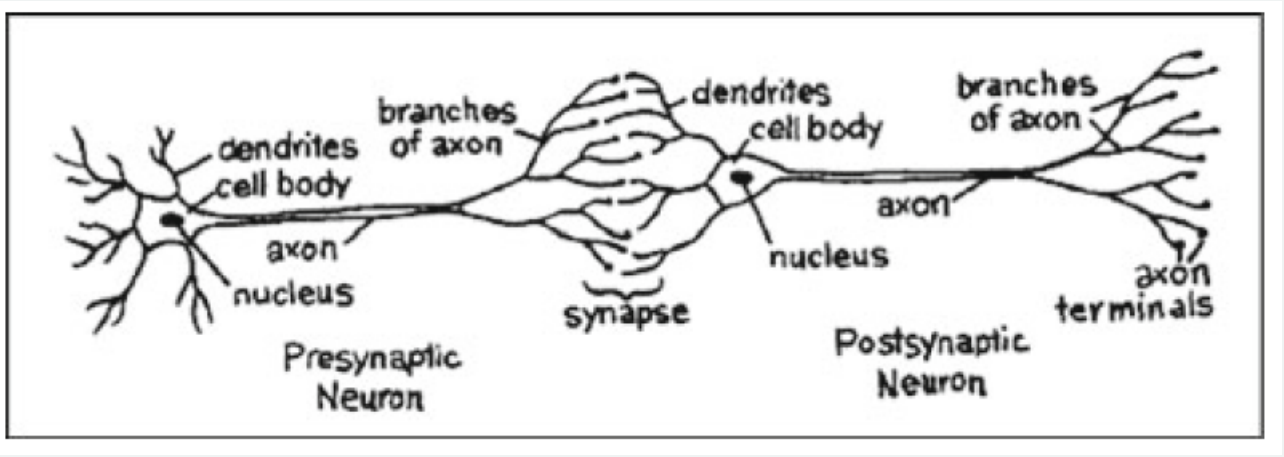
- Arquitectura de red neuronal → **Transformer**
- Muchos datos (tokenizados)
- Miles de GPUs (100.000 Nvidia H100 tiene Elon Musk en su Colossus)
- Fine-tuneado del modelo base para convertirlo en “Instructor”.
- Finalmente, entrenamiento por refuerzo humano

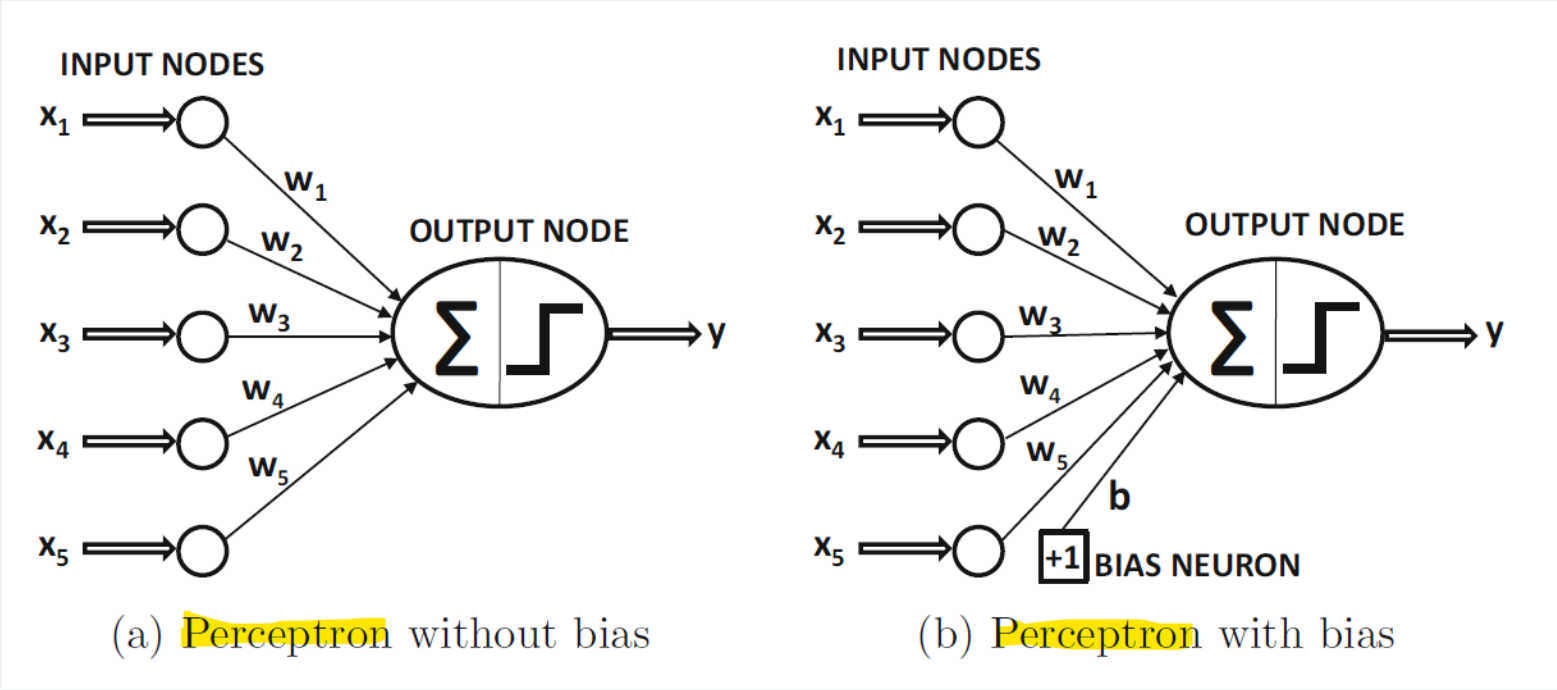




Red neuronal









Antes de continuar...

Tokenizador

Y

Embedding

Tokenizador

- Un **token** en el contexto del procesamiento de lenguaje natural (*NLP)* es una unidad básica de texto que se usa para analizar y procesar el lenguaje. Un *token* puede ser una **palabra**, un **símbolo**, un carácter o una **sub-palabra** (ejemplo: zapat-ero).
- ¿Por qué tokenizar?

➢ Manejo de Palabras Desconocidas (Out-of-Vocabulary)

➢ Reducción del Tamaño del Vocabulario (➔ menos uso de memoria)

➢ Captura de Morfología y Semántica: se identifica una raíz común en “zapatería” y “zapatero”

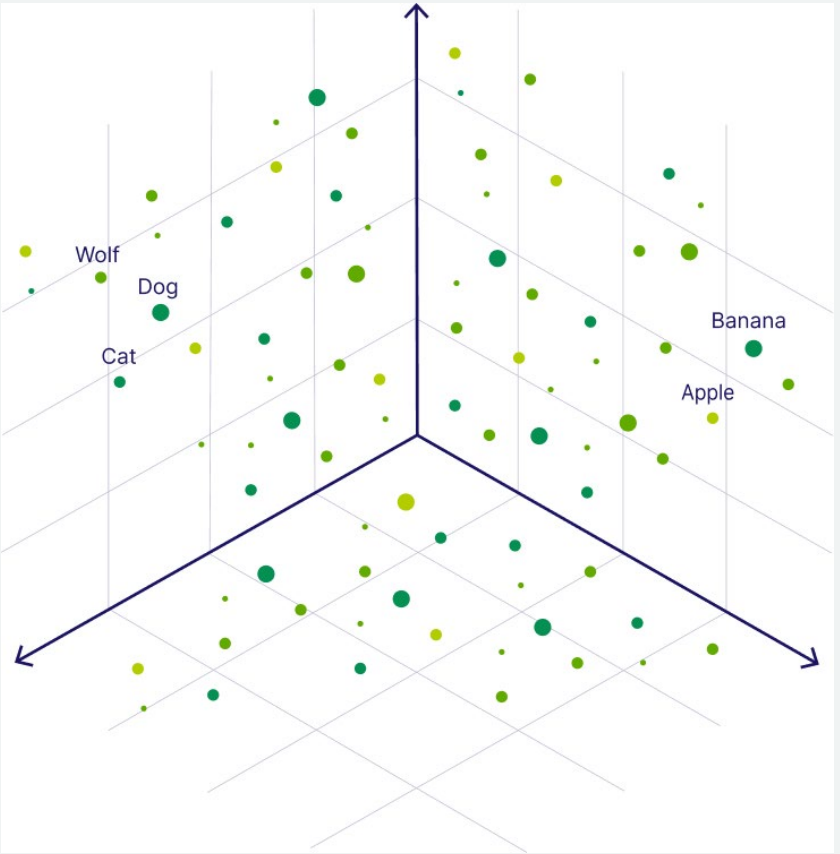


Embeddings: definición

- Un **embedding** es una representación matemática de una palabra (o token) en un espacio multi-dimensional de variables latentes. Convierte palabras en vectores numéricos que capturan la información semántica y sintáctica de esas palabras.

¿Cómo funciona un embedding?

- Los *embeddings* asignan a cada palabra un vector de números reales. Ejemplo: entre 1536 y 3072 en OpenAI.
- Durante el entrenamiento de un modelo de lenguaje, se ajustan los *embeddings* de forma que palabras que tienen significados similares tengan vectores cercanos en este espacio, mientras que palabras con significados diferentes estén más alejadas.




TRANSFORMERS

Transformers

Vaswani et al. (2017)

Elementos clave:

- Mecanismo de atención
- Codificación posicional
- Paralelización



Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

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

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

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].

*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

†Work performed while at Google Brain.

‡Work performed while at Google Research.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

(*) 136.000 citas científicas en 7 años

De forma muy esquemática...¿qué es un LLM (Large Languange Model) en IA?















Dado un texto, genera la **palabra siguiente** (realmente token) **más probable**

(*) ¡Ojo! Tienen fecha de corte de información pero podemos “enchufarles” más contexto



Materiales del curso

Web y necesidad de correo “tipo” gmail

- Los materiales aparecerán en la Web de la UMH:

https://cio.umh.es/cursobenissa

- Además, es necesario disponer de un correo electrónico “tipo” Gmail → crea uno nuevo aquí: https://www.gmail.com/ o usa uno que ya tengas, como quieras.