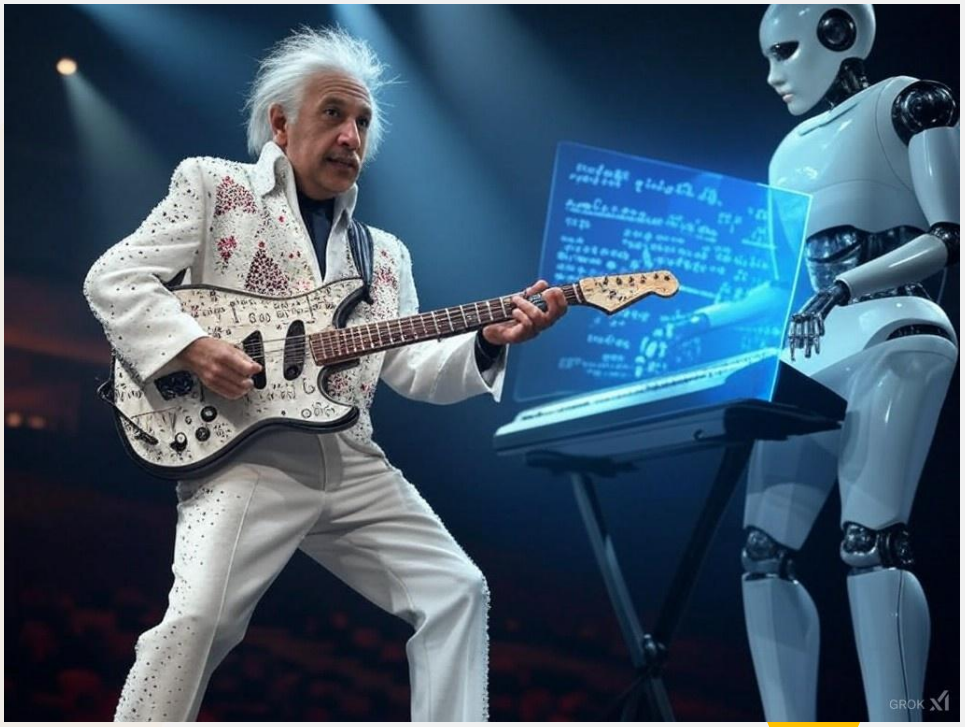
Tema 2: Arquitectura y Funcionamiento de la IA Generativa

Ing. Leopoldo López Gómez MPA MSc
Harvard University



Generado por Grok 2, 2025

CONTENIDO

1. LOS IMPULSORES DEL CRECIMIENTO DE LA IA
2. IA GENERATIVA Y LLMs

SUBTEMA 1
Los impulsores del Crecimiento de la IA Generativa

¿Por qué ahora ?

Avances en la capacidad de procesamiento

Desarrollo de Modelos revolucionarios

Disponibilidad masiva de datos

¿Por qué ahora ?

Avances en la capacidad de procesamiento

Desarrollo de Modelos revolucionarios

Disponibilidad masiva de datos

- Tarjetas gráficas (GPU)
- Inteligencia artificial (IA) y aprendizaje profundo
- Automóviles autónomos
- Data centers y computación en la nube
- Juegos y realidad virtual
- Software y ecosistemas

CEO. Jen - Huan Sun

Nvidia

NASDAQ: NVDA

Resumen | Comparación | Finanzas

Resumen de mercado > Nvidia

144.47 USD

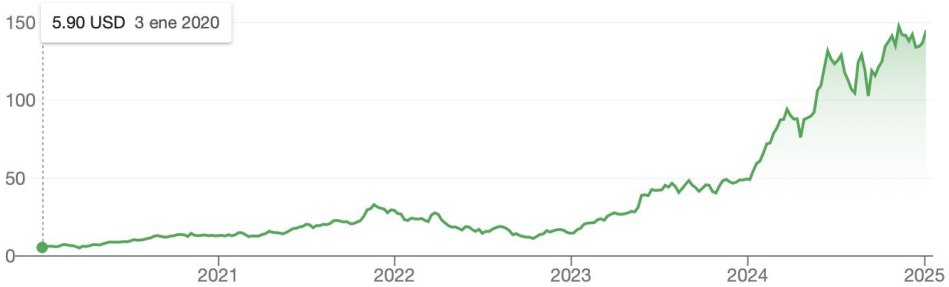
+ Seguir

+138.57 (2,348.64%) ↑ en los últimos 5 años

Cerrado: 3 ene, 8:00 p.m. GMT-5 • Renuncia de responsabilidad

Tras cierre 144.99 +0.52 (0.36%)

1D | 5D | 1 M | 6 M | UAHF | 1A | 5A | Máx.



Abierto	140.01	Cap. burs.	3.54 B	En 52-sem.	152.89
Alta	144.90	Índice PER	56.93	En 52-sem.	47.32
Baja	139.73	Rend. div.	0.028%		

Comentarios

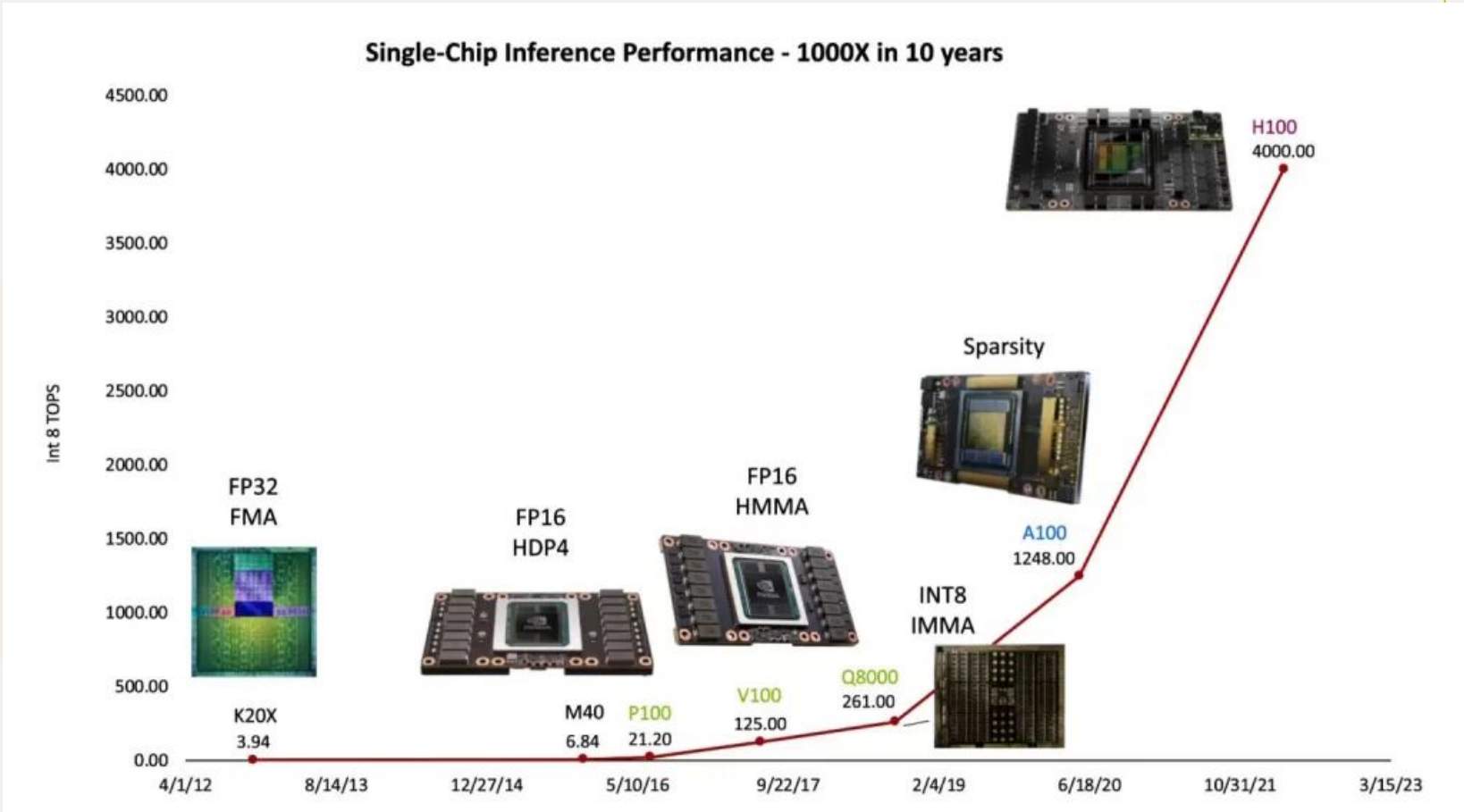
Ver más detalles →

Infraestructura para las IA

- Hardware de alto rendimiento: GPU y TPU
- Almacenamiento masivo
- Red de alta velocidad
- Plataformas de software y herramientas de desarrollo
- Datos de entrenamiento
- Infraestructura de nube

Costo LLM OpenAI ChatGPT-4 : US$100 millones

Infraestructura para las IA



¿Por qué ahora ?

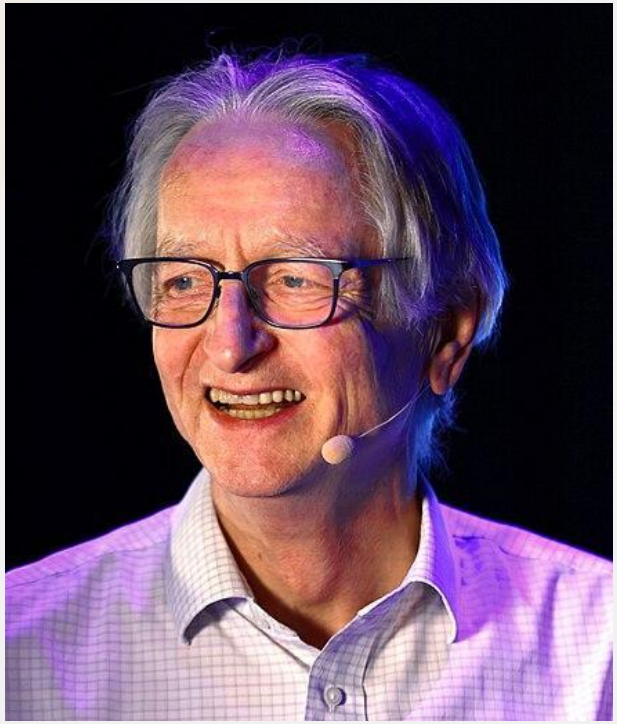
Avances en la capacidad de procesamiento

Desarrollo de Modelos revolucionarios

Disponibilidad masiva de datos

Geoffrey Hinton

- **Padre del Deep Learning**.
- **Backpropagation**: Entrenamiento de redes neuronales.
- **Máquinas de Boltzmann**: Aprendizaje no supervisado.
- **Aplicaciones**: Voz, visión y lenguaje.
- **Premio Turing 2018**.



Geoffrey Hinton

Transformers 2017

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.

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

¿Por qué ahora ?

Avances en la capacidad de procesamiento

Desarrollo de Modelos revolucionarios

Disponibilidad masiva de datos

Disponibilidad Masiva de Datos

- **Crecimiento exponencial de los datos globales**
- **Acceso a fuentes abiertas de datos**
- **Desarrollo de infraestructuras de almacenamiento y procesamiento**
- **Calidad de los datos**
- **Datos en tiempo real**

SUBTEMA 2
IA GENERATIVA Y LLMs

¿Cómo funcionan las IA generativas?

¿Qué son las IA generativas?

Las IA generativas (como ChatGPT) son modelos de lenguaje entrenados con enormes cantidades de datos para generar texto, imágenes, etcétera, de forma automática

¿Cómo funcionan?

Utilizan redes neuronales profundas entrenadas con el aprendizaje por refuerzo para predecir la siguiente palabra o pixel basado en los datos de entrenamiento

Ventajas y limitaciones

Facilitar tareas repetitivas, así como emular procesos creativos

La información puede ser sesgada o incorrecta

Deben usarse éticamente

Las IA generativas como ChatGPT prometen revolucionar las tareas creativas, pero deben usarse responsablemente.

Características de las IA generativas como ChatGPT

Arquitectura *Transformer*

Procesamiento paralelo eficiente y captura de relaciones de largo alcance en secuencias de texto que facilita la comprensión y generación de texto coherente y gramaticalmente correcto

Preentrenamiento y afinamiento

Se preentrena en grandes conjuntos de datos para aprender patrones y estructuras del lenguaje

Se afina en conjuntos más pequeños para adaptarse a tareas como responder preguntas o mantener conversaciones

Generación de texto

ChatGPT genera texto dinámica y creativamente Conversa más naturalmente y genera contenido de manera similar como lo hacen las personas

Adaptabilidad

A una amplia gama de tareas y dominios dentro del procesamiento del lenguaje natural, como traducción automática, resumen de textos, generación de contenido, preguntas y respuestas, y más

Contextualización

Con la arquitectura *Transformer* puede tener en cuenta el contexto del texto de entrada para proporcionar respuestas más precisas y relevantes

Manejo de múltiples idiomas

Entrena principalmente en inglés, pero también puede manejar otros idiomas según datos disponibles. Incluye múltiples lenguajes de programación


Modelos de Lenguaje

Empresa	Modelo de Lenguaje	No. usuarios millones	Fecha
Google	Gemini 2.5 Pro	450	Junio 2025
Amazon (Anthropic)	Claude Opus 4	50	Mayo 2025
Microsoft (OpenAI)	ChatGPT 5	800	7 ago 2025
Meta (Facebook)	LLaMA 4	300	Abril 2025
Baidu	ERNIE 4.5 Turbo	300	Abril 2025
Alibaba	Qwen 3	150	Abril 2025
Tencent	Hunyuan 2.5	20 - 35	Abril 2025
DeepSeek	DeepSeek v3.1	100	21 Ago 2025

Foundational Large Language Models

ChatGPT

Claude

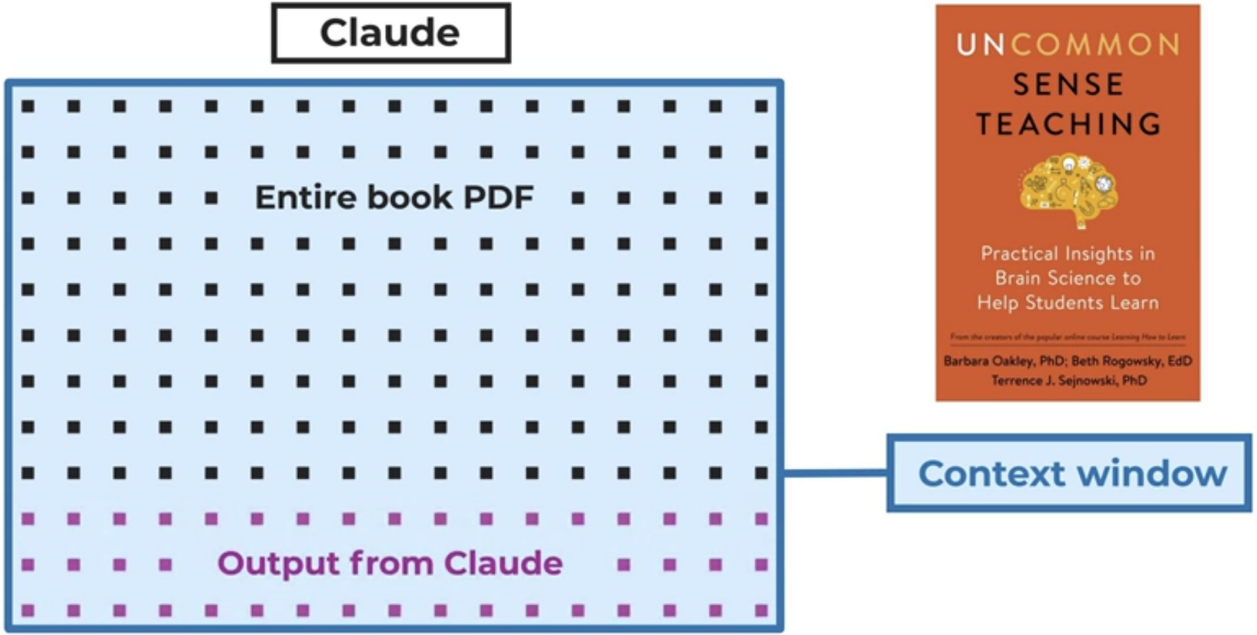
Gemini

7 International Giants with their Foundational Large Language Models ("Engines")

Google	Amazon	Microsoft	(Facebook /Meta)	Baidu	Alibaba	Tencent
Gemini	ANTHROPIC Claude 3.5 Sonnet	OpenAI ChatGPT 4.0	Meta LLaMA 3	"ERNIE" 文心一言	Tongyi Qianwen 通义	Hunyuan 腾讯混元
1 million tokens	200,000 tokens	128,000 tokens	70,000 tokens	"competitive"	"competitive"	"competitive"

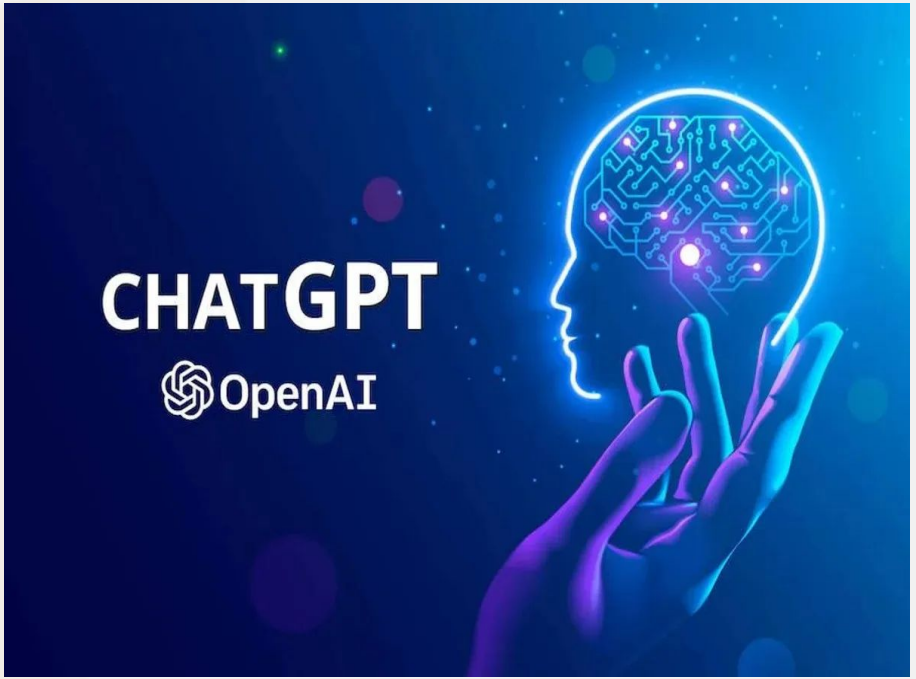
Context window

Words you are trying to put in + words you want to get out

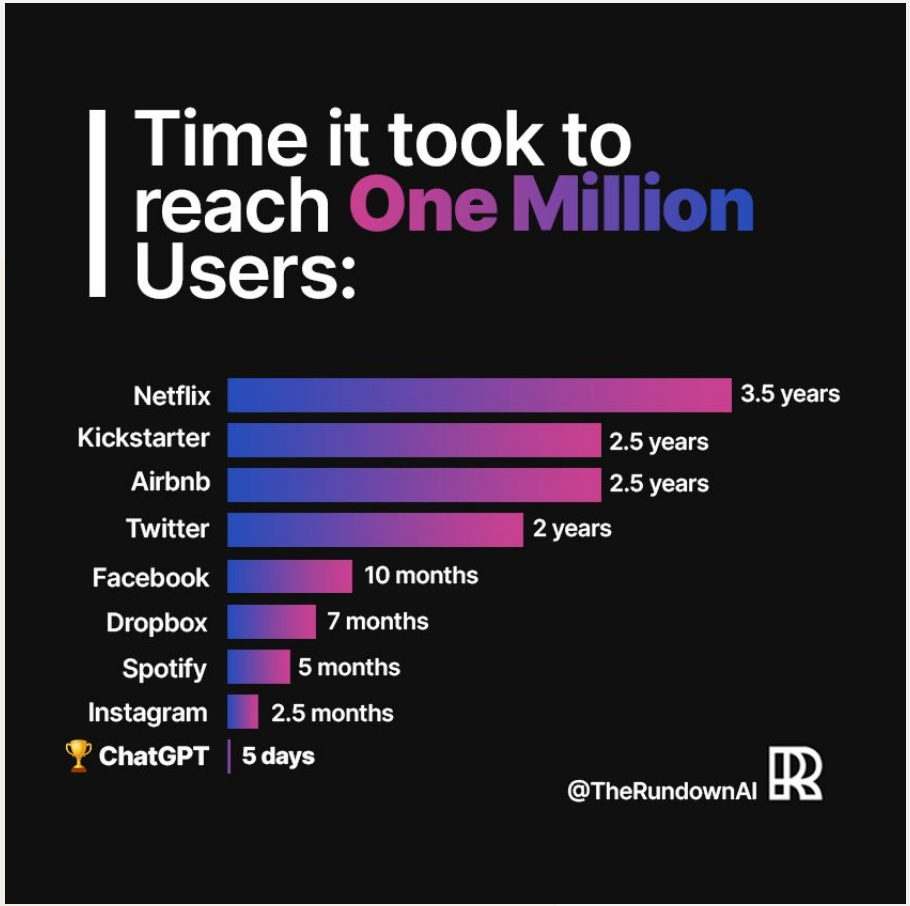


Chat GPT

- Generación de Texto y Respuestas
- Resumir Texto
- Traducción de más de 100 Idiomas
- Escuchar (Reconocimiento de Voz)
- Ver (Procesamiento de Imágenes)
- Conversación Continua y Contextual
- Generación de Resúmenes y Síntesis
- Personalización y Adaptabilidad



Time to Reach 1 Million users



Llegó GPT-5

Nuestro modelo más inteligente, rápido y útil hasta ahora.

Gemini

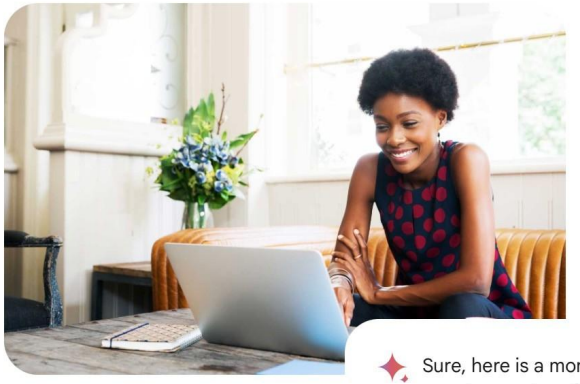
For business FAQ

Gemini

Supercharge your creativity and productivity

Chat to start writing, planning, learning and more with Google AI

Chat with Gemini



Sure, here is a more clear and concise version of your email draft:

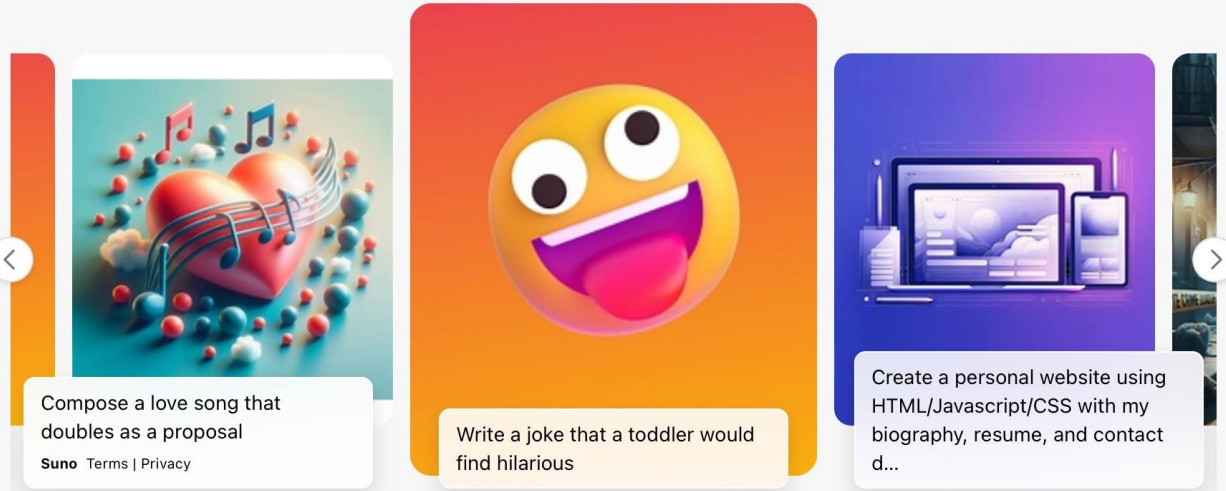
Dear Professor [Professor's name],

Congratulations on your Teaching

Google Privacy & Terms

Copilot

Your everyday AI companion



Ask me anything...

0/2000

Claude

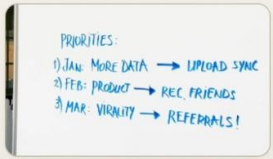
Good afternoon, Jordi

What can I help you with?

Start Chat

New in Claude

Understand and work with images



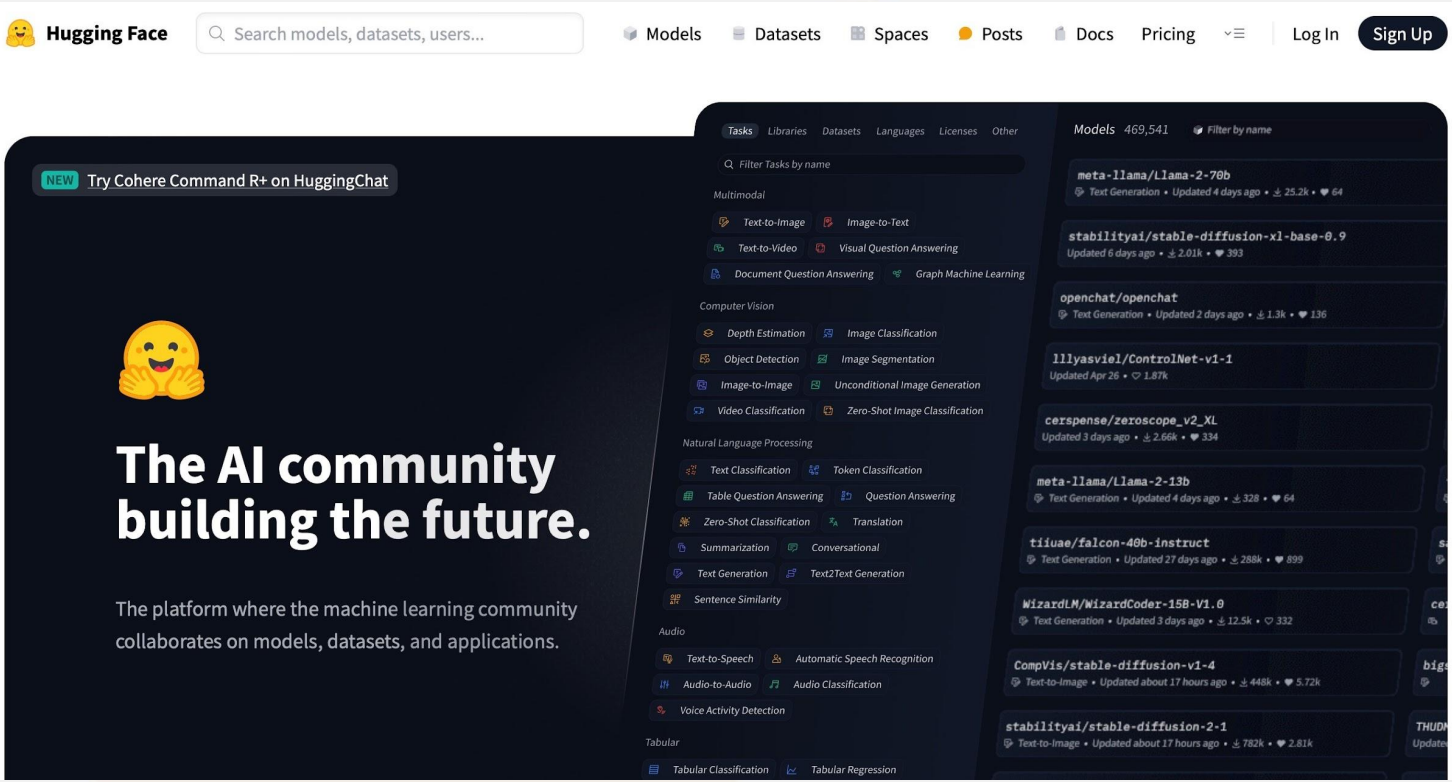
Transcribe handwritten notes

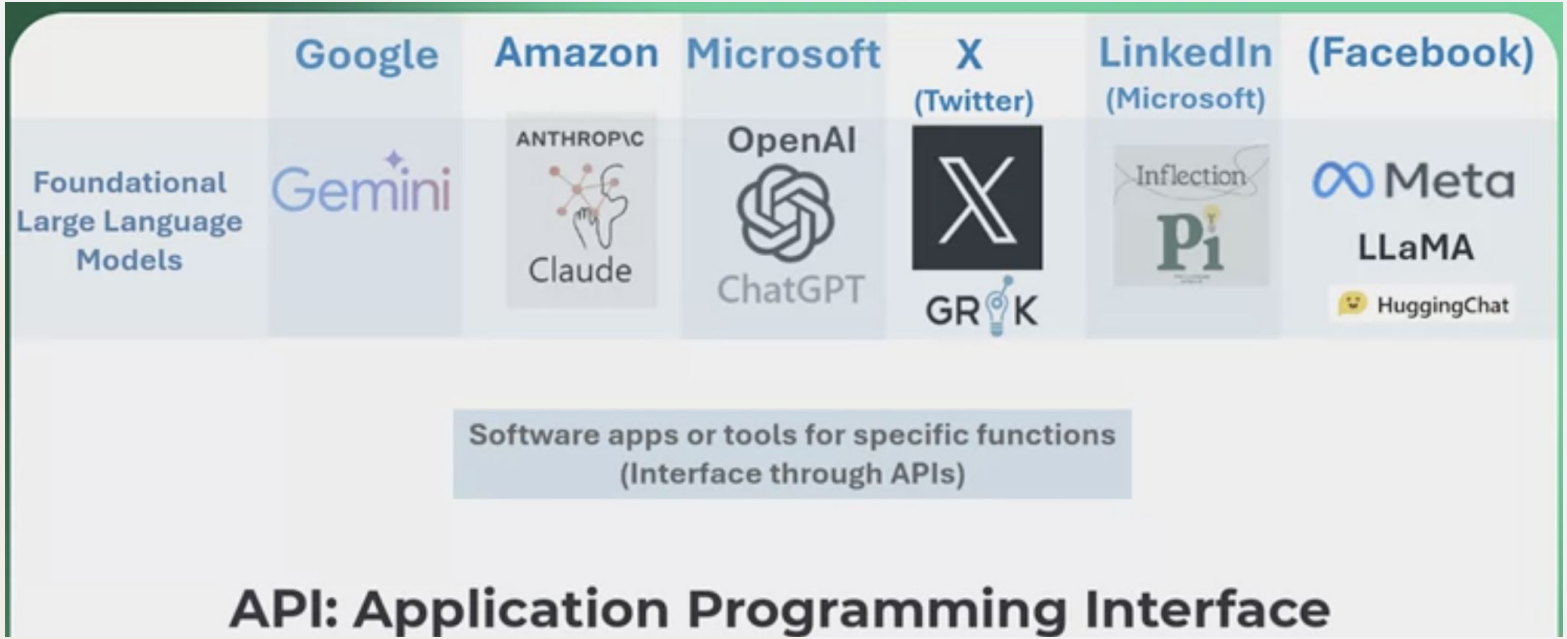
Extract text from images

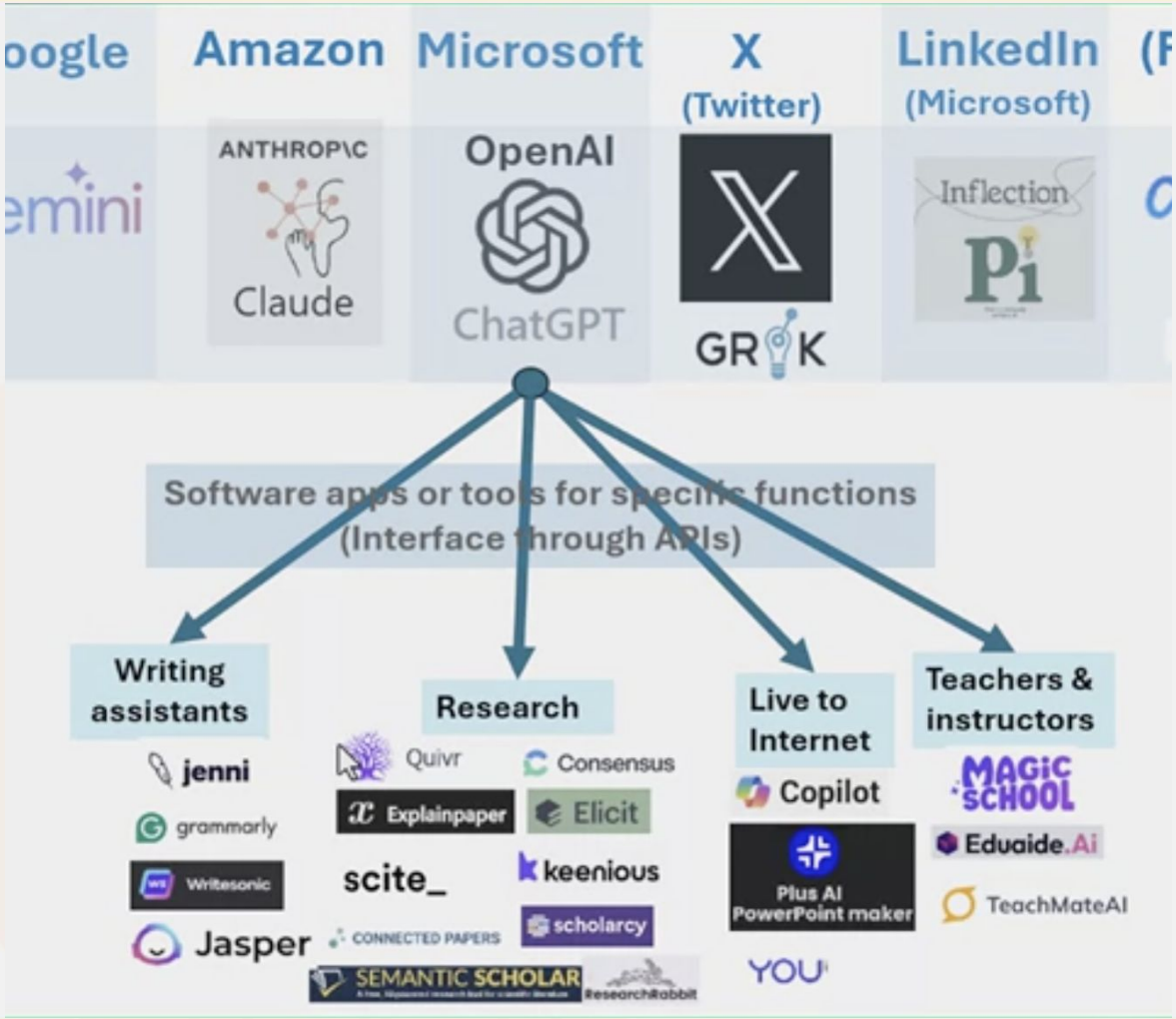
Recommend style improvements

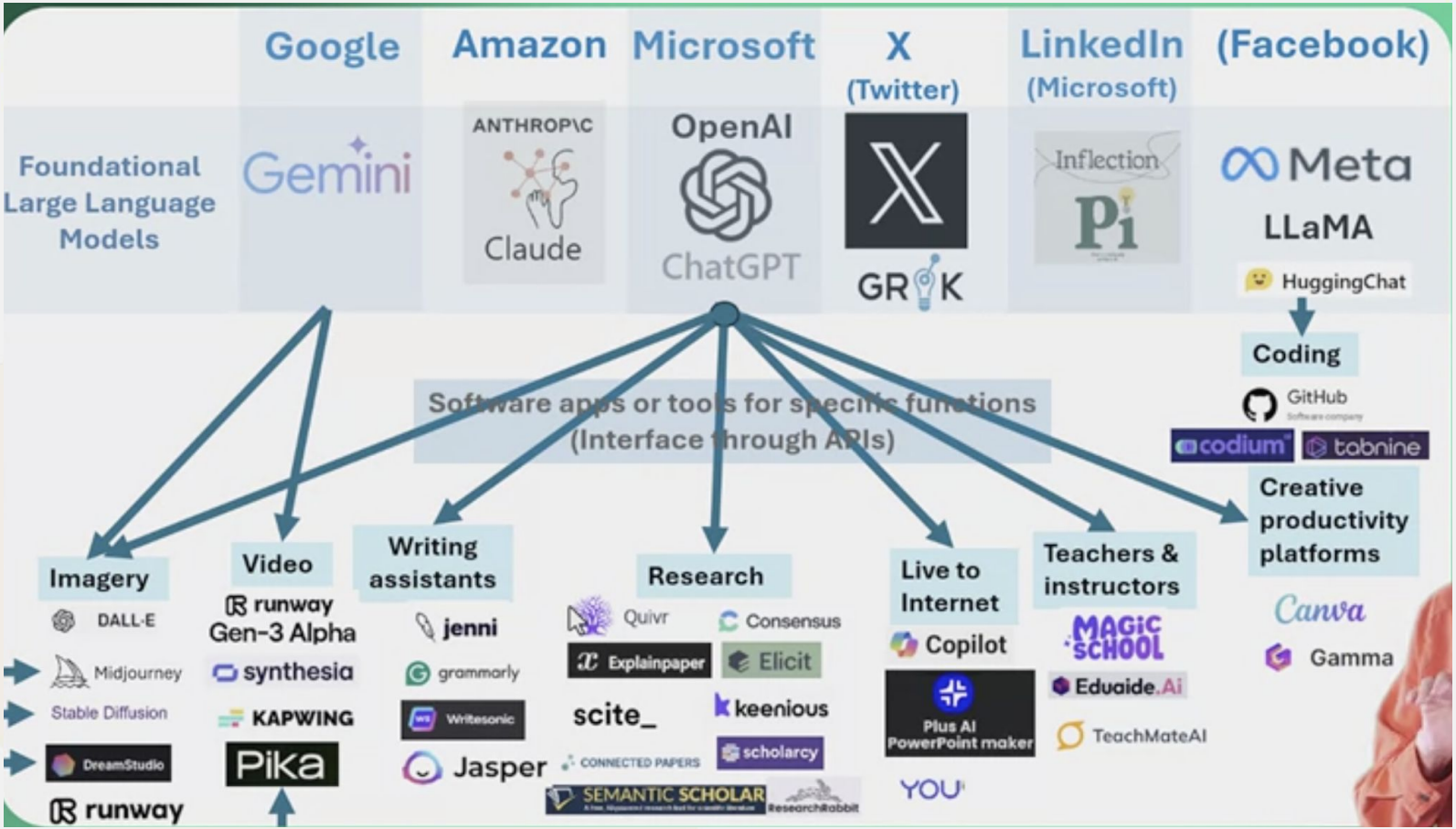
Previous chats from 3 days ago

ANTHROP\C


Hugging Face
Search models, datasets, users...
Models
Datasets
Spaces
Posts
Docs
Pricing
Log In
Sign Up
NEW Try Cohere Command R+ on HuggingChat
The AI community building the future.
The platform where the machine learning community collaborates on models, datasets, and applications.
Tasks
Libraries
Datasets
Languages
Licenses
Other
Filter Tasks by name
Multimodal
Text-to-Image
Image-to-Text
Text-to-Video
Visual Question Answering
Document Question Answering
Graph Machine Learning
Computer Vision
Depth Estimation
Image Classification
Object Detection
Image Segmentation
Image-to-Image
Unconditional Image Generation
Video Classification
Zero-Shot Image Classification
Natural Language Processing
Text Classification
Token Classification
Table Question Answering
Question Answering
Zero-Shot Classification
Translation
Summarization
Conversational
Text Generation
Text2Text Generation
Sentence Similarity
Audio
Text-to-Speech
Automatic Speech Recognition
Audio-to-Audio
Audio Classification
Voice Activity Detection
Tabular
Tabular Classification
Tabular Regression
Models 469,541
Filter by name
meta-llama/Llama-2-70b
Text Generation • Updated 4 days ago • 25.2k • 64
stabilityai/stable-diffusion-xl-base-0.9
Updated 6 days ago • 2.01k • 393
openchat/openchat
Text Generation • Updated 2 days ago • 1.3k • 136
lllyasviel/ControlNet-v1-1
Updated Apr 26 • 1.87k
cerspense/zeroscope_v2_XL
Updated 3 days ago • 2.66k • 334
meta-llama/Llama-2-13b
Text Generation • Updated 4 days ago • 328 • 64
tiiuae/falcon-40b-instruct
Text Generation • Updated 27 days ago • 288k • 899
WizardLM/WizardCoder-15B-V1.0
Text Generation • Updated 3 days ago • 12.5k • 332
CompVis/stable-diffusion-v1-4
Text-to-Image • Updated about 17 hours ago • 448k • 5.72k
stabilityai/stable-diffusion-2-1
Text-to-Image • Updated about 17 hours ago • 782k • 2.81k


Google
Amazon
Microsoft
X
(Twitter)
LinkedIn
(Microsoft)
(Facebook)
Foundational Large Language Models
Gemini
ANTHROPIC
Claude
OpenAI
ChatGPT
GROK
Inflection
Pi
Meta
LLaMA
HuggingChat
Software apps or tools for specific functions
(Interface through APIs)
API: Application Programming Interface


oogle
Amazon
Microsoft
X
(Twitter)
LinkedIn
(Microsoft)
emini
ANTHROP\C
Claude
OpenAI
ChatGPT
GROK
Inflection
Pi
Software apps or tools for specific functions
(Interface through APIs)
Writing assistants
jenni
grammarly
Writesonic
Jasper
Research
Quivr
Consensus
Explainpaper
Elicit
scite_
keenious
scholarcy
CONNECTED PAPERS
SEMANTIC SCHOLAR
ResearchRabbit
Live to Internet
Copilot
Plus AI PowerPoint maker
YOU
Teachers & instructors
MAGIC SCHOOL
Eduaide.Ai
TeachMateAI


Google
Amazon
Microsoft
X (Twitter)
LinkedIn (Microsoft)
(Facebook)
Foundational Large Language Models
Gemini
ANTHROPIC
Claude
OpenAI
ChatGPT
GROK
Inflection
Pi
Meta
LLaMA
HuggingChat
Coding
GitHub
codium
tabnine
Software apps or tools for specific functions
(Interface through APIs)
Creative productivity platforms
Canva
Gamma
Imagery
DALL·E
Midjourney
Stable Diffusion
DreamStudio
runway
Video
runway Gen-3 Alpha
synthesia
KAPWING
Pika
Writing assistants
jenni
grammarly
Writesonic
Jasper
Research
Quivr
Consensus
Explainpaper
Elicit
scite_
keenious
scholarcy
CONNECTED PAPERS
SEMANTIC SCHOLAR
ResearchRabbit
Live to Internet
Copilot
Plus AI PowerPoint maker
YOU
Teachers & instructors
MAGIC SCHOOL
Eduaide.Ai
TeachMateAI

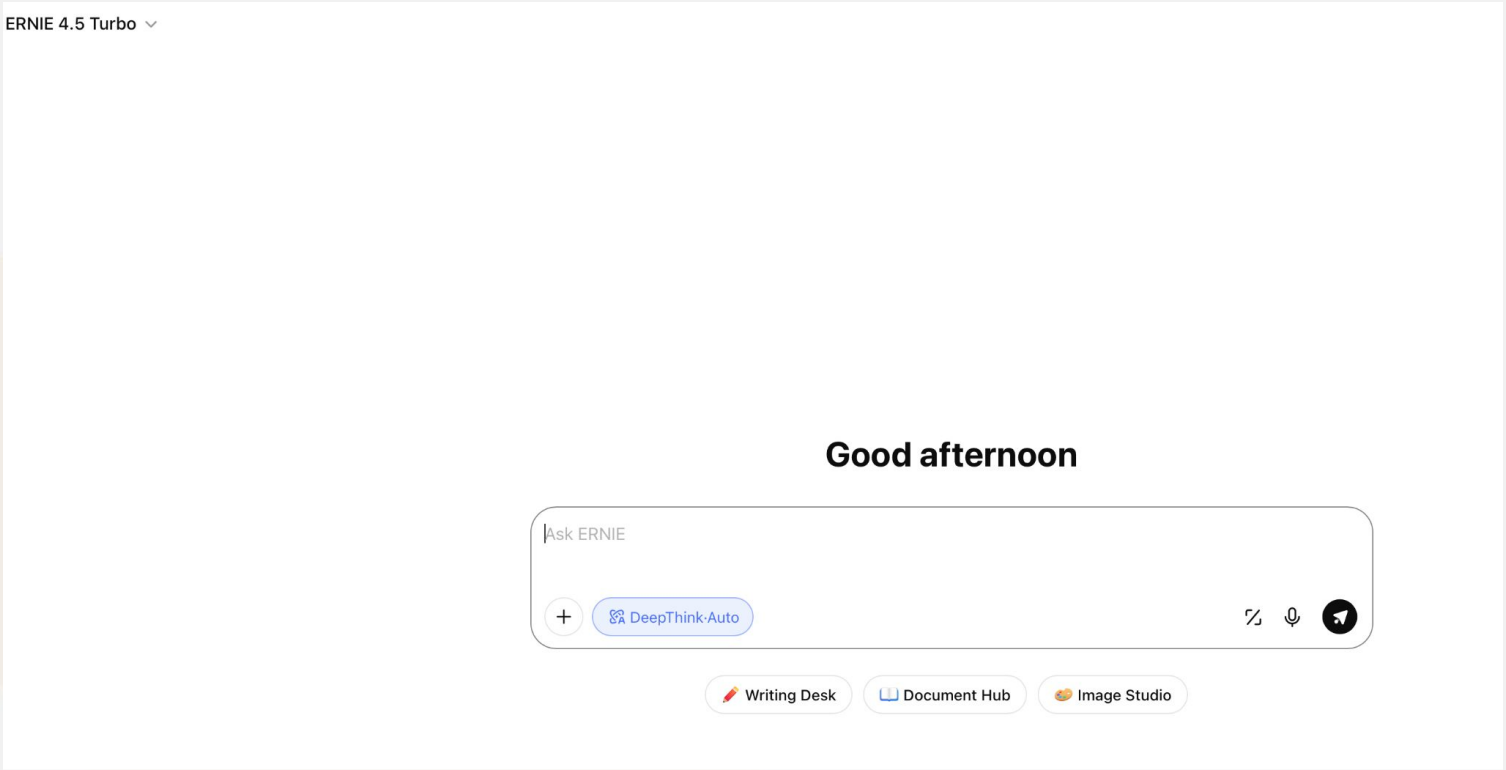
US vs China

Empresas chinas relacionadas con la IA suman 1.67 millones en el primer semestre de 2024. Sumaron más de 237,000 nuevas empresas involucradas en IA durante en ese periodo

POLICY / ARTIFICIAL INTELLIGENCE / TECH

Nvidia's H800 AI chip for China is blocked by new export rules

Baidu / Ernie 4.5 Turbo



https://ernie.baidu.com/

Alibaba / Qwen 3.0

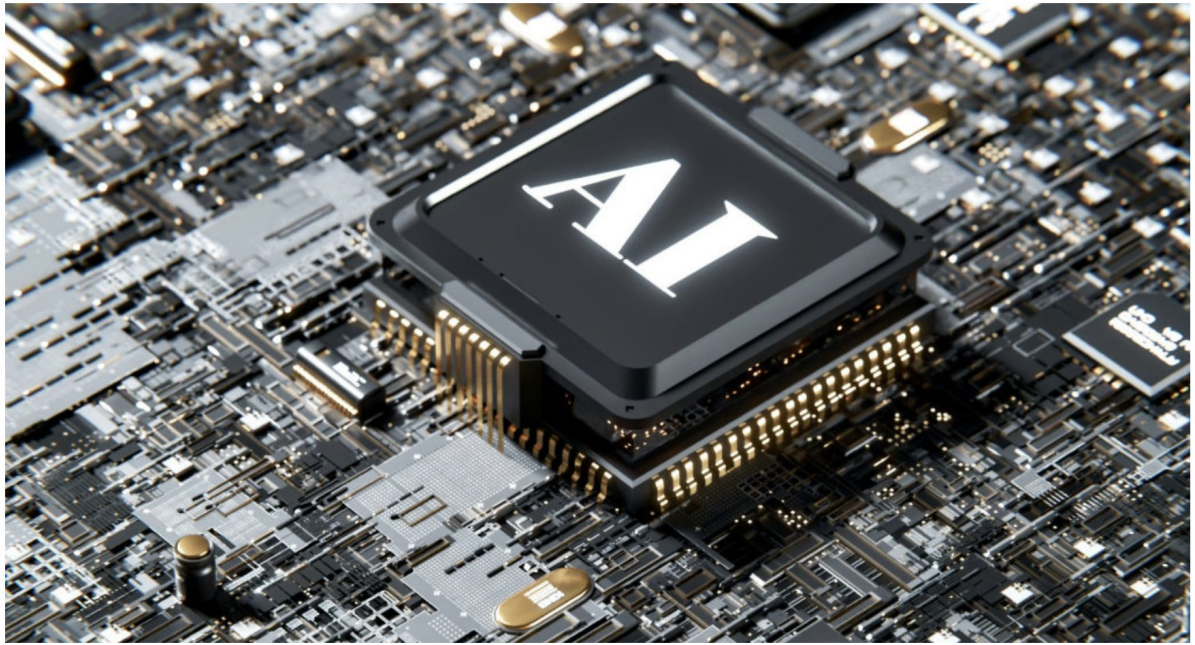
https://chat.qwen.ai/

Tiemblan ChatGPT y Gemini: un chatbot chino con inteligencia artificial sacude Silicon Valley

Se trata de Deepseek V3, que generó un gran revuelo en el sector tecnológico, así como también en la bolsa de Wall Street.

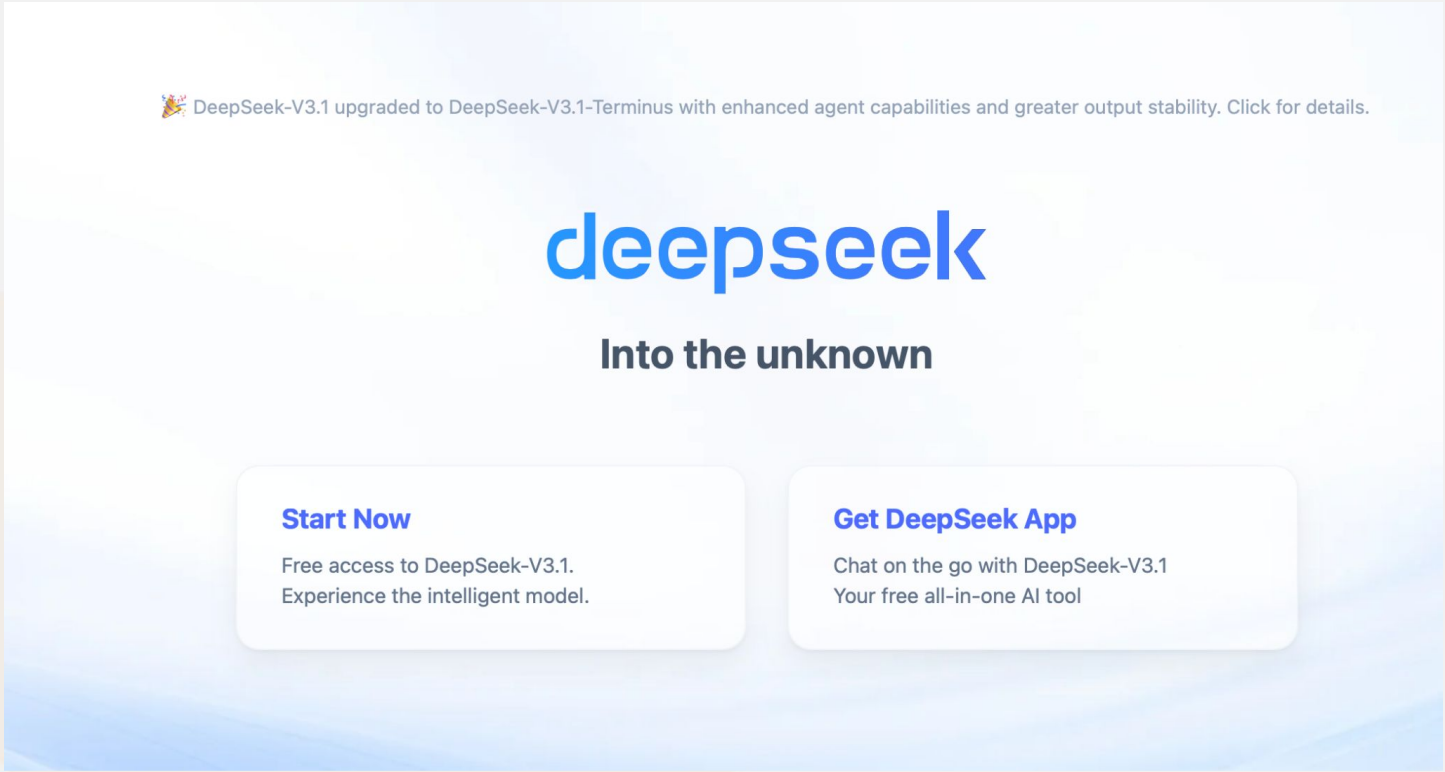
Por Canal26

Viernes 3 de Enero de 2025 - 13:18

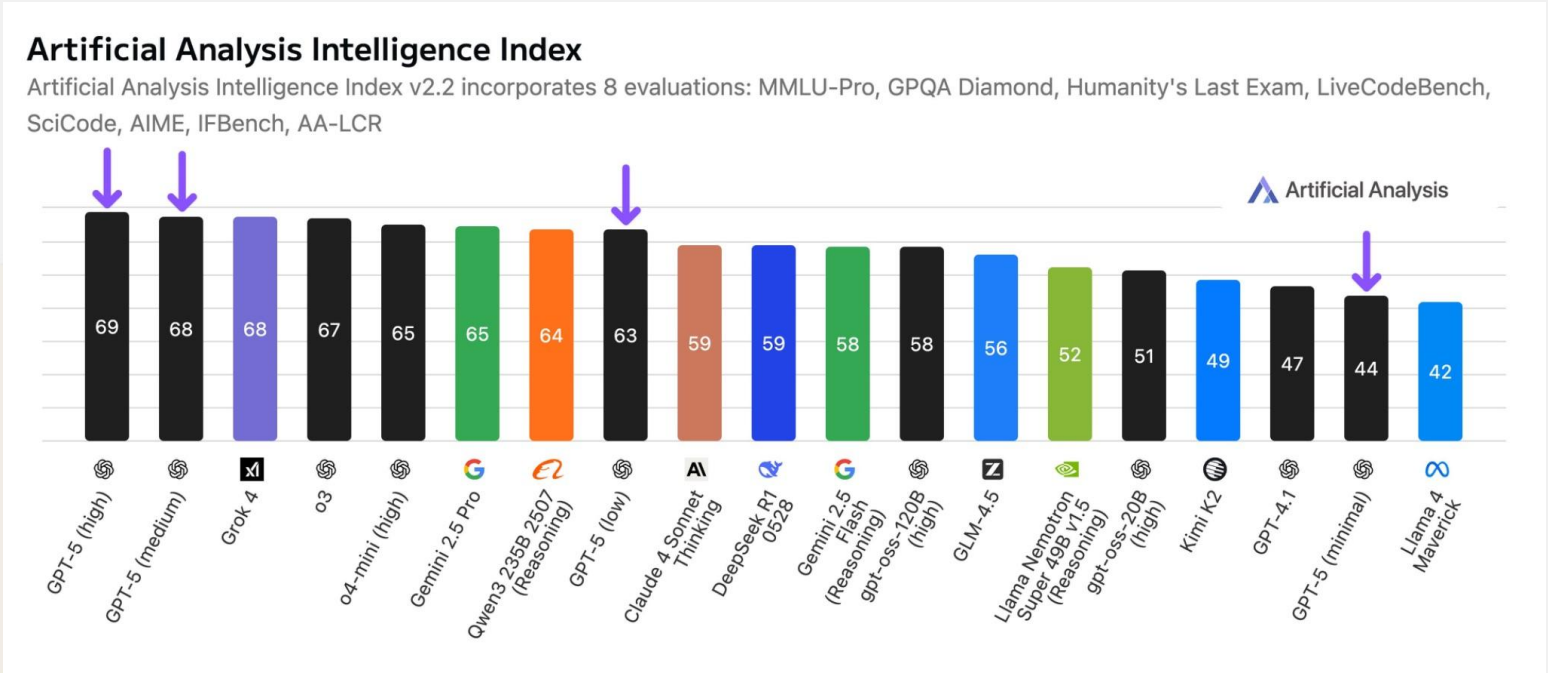


Inteligencia Artificial. Foto: Usplash

Deepseek 3.1.



https://www.deepseek.com



NATHAN LAMBERT

AUG 07, 2025

Top 5 mundial de modelos LLM – octubre 2025

Posición	Modelo	Elo global	Codificación	Visión	AAII v3	MMLU-Pro (%)	ARC-AGI v2
1	Gemini 2.5 Pro (Google)	1466	1469	1266	63	86.2	4.9
2	Grok-4-0709 (xAI)	1446	1453	1221	61	85.4	4.6
3	GPT-5 (OpenAI)	1443	1462	1248	62	85.8	4.8
4	Claude Sonnet 4.5 (Anthropic)	1431	1441	1212	60	84.9	4.5
5	Qwen 2.5 Max (Alibaba Cloud)	1409	1433	1207	58	83.7	4.3

¿Qué modelo de IA usar según la tarea?

Modelo	Fortaleza Principal	Benchmark Clave	Puntuación Destacada	Ideal para...
Gemini 2.5 Pro	Análisis Multimodal (texto + imagen)	MMMU	79.6%	Analizar documentos con gráficos, auditorías visuales, investigación científica.
GPT-5	Programación Algorítmica	HumanEval	92.7%	Desarrolladores, resolución de problemas de código, integración en ecosistema Microsoft.
Claude 4.5	Seguridad y Codificación Real	SWE-bench	72.5%	Proyectos empresariales, mantenimiento de código, entornos con altos requisitos de seguridad.
Grok-4	Contexto Conversacional	DialogQA	84.3%	Atención al cliente avanzada, análisis de diálogos largos, coherencia narrativa.


Cómo funcionan los LLM?

IA Generativa

IA generativa

- Objetivo: Crear nueva información
- Proceso: Expandir tu información

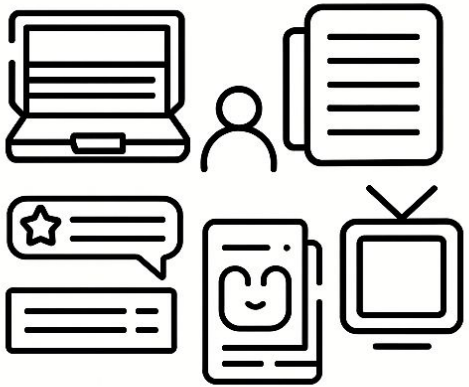
Castigo divino

Sherlock Holmes, tras seguir las huellas del veneno y la corrupción en León, descubrió que la muerte de Cabrera no fue obra de un castigo divino, sino de la ambición humana, sellada en una copa de vino y un pacto de silencio

Crear una historia corta de Sherlock Holmes resolviendo el caso de Castigo Divino de Sergio Ramírez

ADQUISICIÓN DE LENGUAJE

Maximizar la exposición a materiales



Modelos de preentrenamiento del lenguaje

- Usando textos de archivos de internet
- La “P” en “GPT” significa “Preentrenamiento”

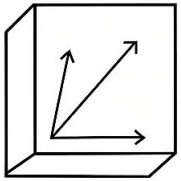
AI Generativa = Modelos matemáticos

IA generativa

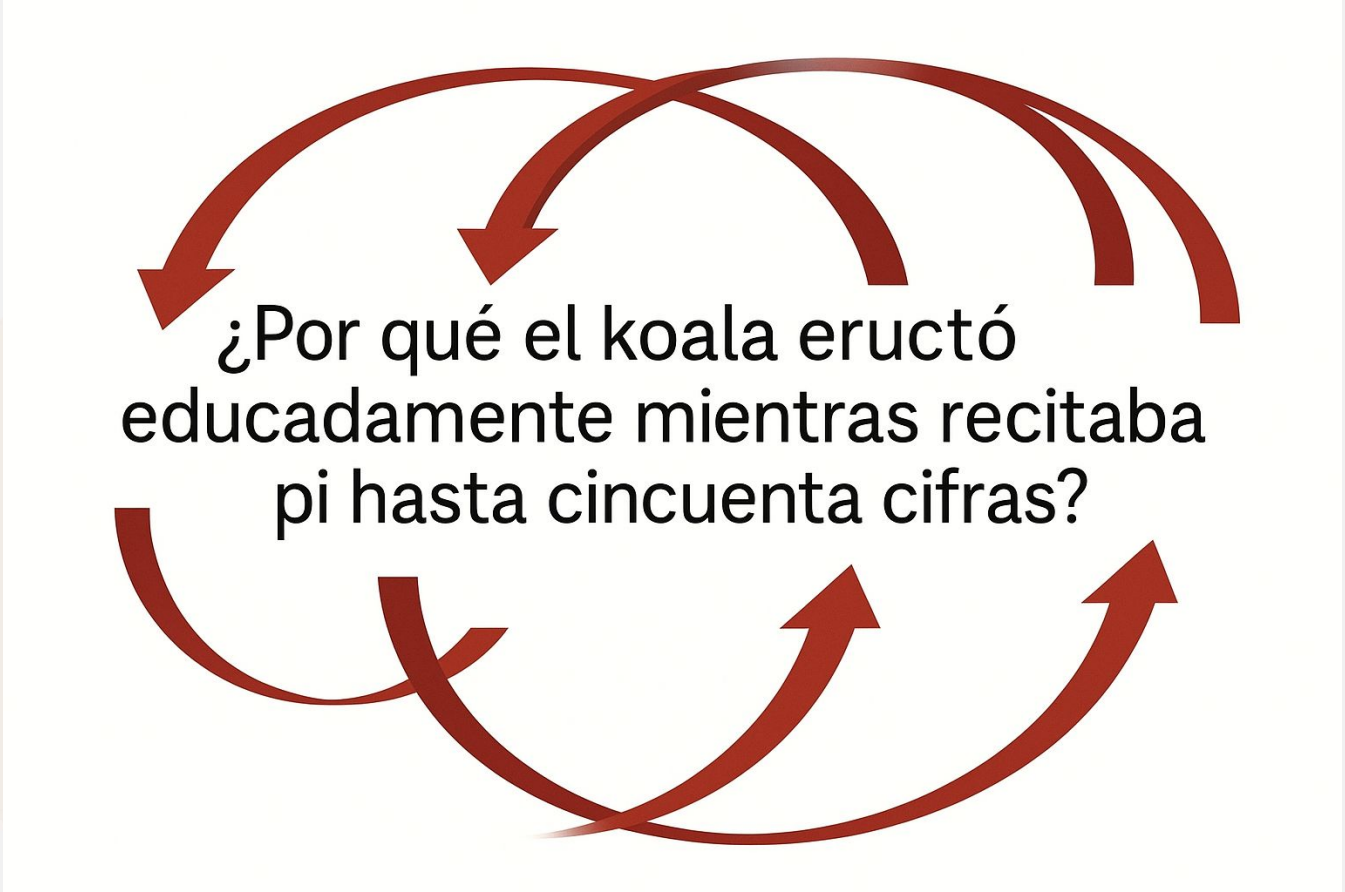
- Objetivo: Crear nueva información
- Proceso: Expandir tu información

Información generada

- Lenguaje (texto)
- Imagen
- Video
- Objetos 3D
- Código



Puede representarse como **vectores**


¿Por qué el koala eructó
educadamente mientras recitaba
pi hasta cincuenta cifras?

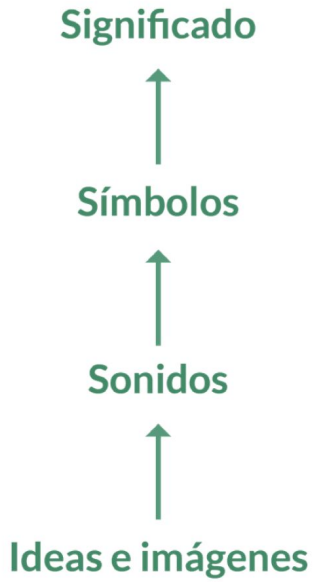
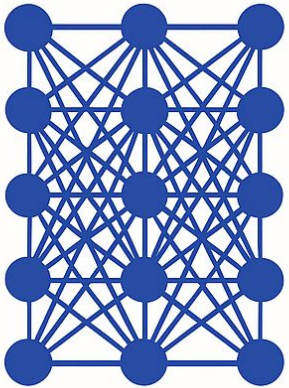
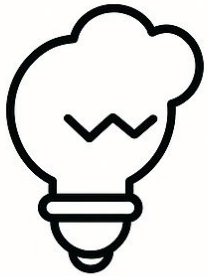


Figura 2: Componentes del lenguaje *(elaboración propia)*

Creación de lenguaje

Después del pre-entrenamiento del LLM





1. Prompt

Qué quieres que diga el modelo?

2. Ensamblaje

Predicción recursiva de las siguientes palabras

Qué es un Token?

Un **token** puede ser una palabra, una parte de una palabra (como sílabas o prefijos/sufijos), o incluso un símbolo (como un signo de puntuación).

- Tokenización
- Procesamiento
- Límite de Tokens (8,192 en GPT-4)
- Costo por token (API)
- **$0.03 por 1,000 tokens** (aproximadamente 750 palabras).

Tokenización (Bag of Words)

Untitled1.ipynb

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda

Comandos + Código + Texto Ejecutar todo

11 s

```
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer

corpus = ["La IA está cambiando la educación", "La tokenización convierte frases en tokens"]

vectorizer = CountVectorizer()
X = vectorizer.fit_transform(corpus)

print(vectorizer.get_feature_names_out())
print(X.toarray())
```

```
['cambiando' 'convierte' 'educación' 'en' 'está' 'frases' 'ia' 'la'
 'tokenización' 'tokens']
[[1 0 1 0 1 0 1 2 0 0]
 [0 1 0 1 0 1 0 1 1 1]]
```

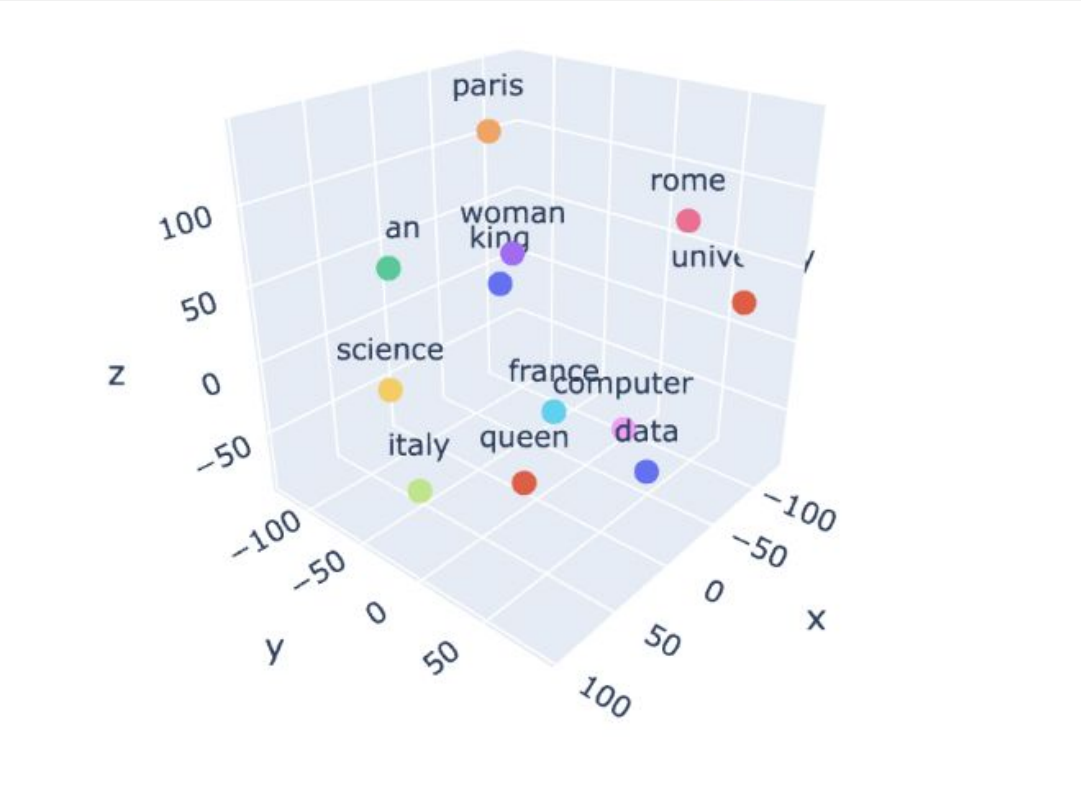
Tokenización (Word Embeddings)

```python
# Obtener el vector completo de la palabra "king"
vector_king = model.wv["king"]

print("Dimensiones:", vector_king.shape)
print("Vector completo:\n", vector_king)
```

```
Dimensiones: (100,)
Vector completo:
 [ 0.0137697   0.4061347  -0.12626676 -0.15600555 -0.17770834  0.07054882
  0.10727259  0.15129381  0.06816257  0.19405758 -0.11549628  0.05367924
 -0.19880046 -0.07495552 -0.07015217  0.02174423  0.20768861 -0.3095207
 -0.19560754 -0.12425077  0.15505514  0.09234575  0.37904513 -0.13223383
 -0.19674362 -0.19474052 -0.57882583 -0.20982927  0.16001005  0.14618084
  0.28559306 -0.05990702  0.16890408 -0.21574426  0.2558106   0.15756333
  0.26020205  0.18257251 -0.21412925 -0.10700825  0.14373171 -0.30436245
 -0.05264387 -0.03954915  0.27635667  0.14004181 -0.28308302  0.07765327
  0.2746572   0.06703892  0.33313638 -0.40691543 -0.21277665 -0.05297504
 -0.25584447  0.09648985  0.16029395 -0.13740887 -0.08167733  0.22274956
 -0.00317478 -0.00154242  0.01319763 -0.03363317  0.1765131   0.25609913
  0.12346422  0.35898736 -0.2292304   0.19702204 -0.03634495  0.2273616
  0.10911318  0.3838134  -0.06410102 -0.08823231 -0.1733891  -0.15408166
  0.02660611 -0.05687038  0.00650771 -0.14969343 -0.02891455  0.17507587
 -0.05688329 -0.244691    0.27969143  0.00911525  0.08939365  0.26356953
  0.2871455  -0.15651086  0.0202577  -0.224132    0.1814691   0.12413165
  0.31939876  0.25938693 -0.21804802  0.02313315]
```

Visualización 3D interactiva de Word2Vec



LM Visualization

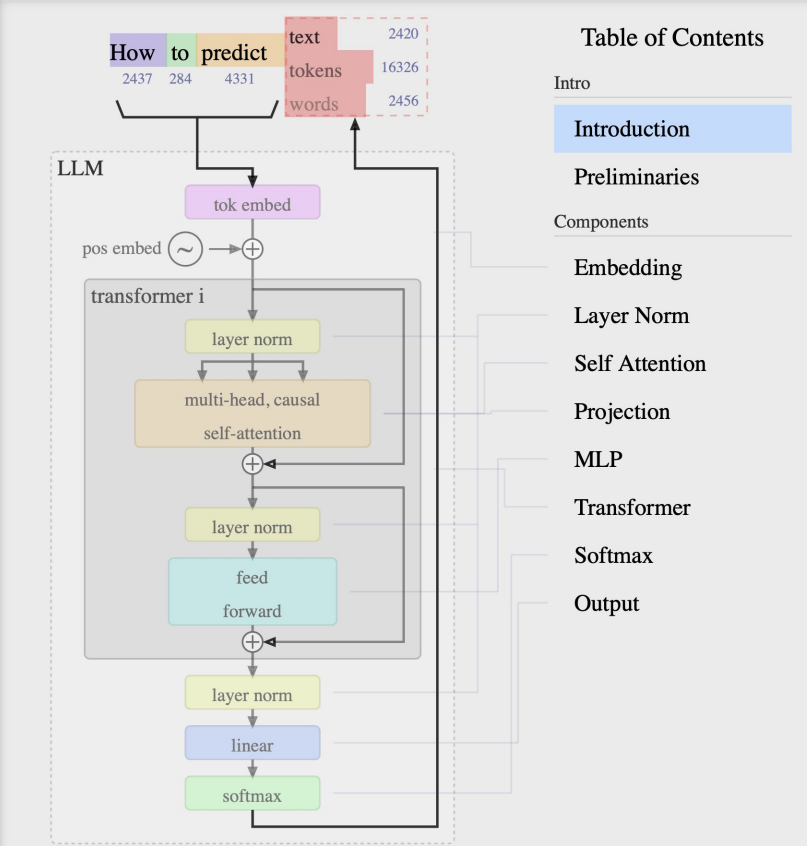
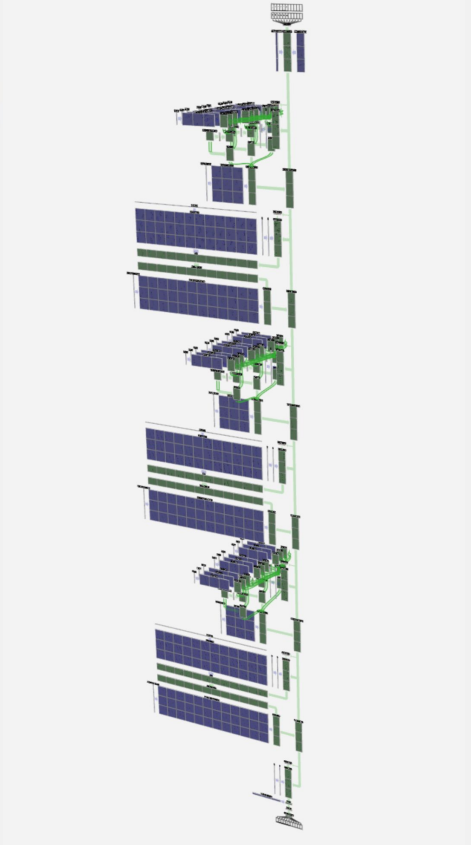


Table of Contents



https://bbycroft.net/llm

Curso IA - Acceso Libre

IA EN BUSINESS

IA PARA LA INNOVACION - ACCESO LIBRE

Leopoldo Lopez

0 de 62 actividades completadas

0% Curso completado

Ver curso

¡Escanéame!