



Aplicaciones de modelos  
generativos avanzados

# ChatGPT

Ramón García Verjaga  
José Alberto Gómez García

# Índice I

## Introducción

Contexto  
¿Qué vamos a tratar  
durante la presentación?

01

## Aplicaciones

03

## ChatGPT

¿Qué es?  
Limitaciones  
Arquitectura  
Técnicas

02

Prompt  
engineering

04

# Índice II

## Plugins

Nuevas  
funcionalidades de  
ChatGPT

05

¿Clasificar  
texto generado  
por AI?  
¿AI sí, AI no?

06

## Conclusiones

07

## Bibliografía

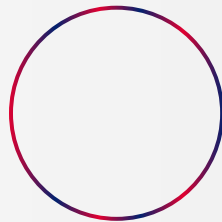
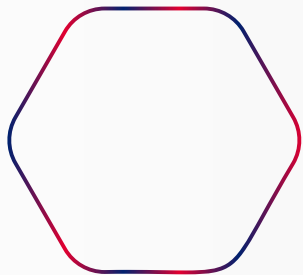
Fuentes de  
información  
utilizadas

08



# 01

## Introducción



# 02

## ChatGPT

# ChatGPT | ¿Qué es?

## *Generative Pre-trained Transformer, GPT*

- es un **modelo** basado en la **arquitectura transformer** que ha sido **preentrenado** gracias a grandes cantidades de **texto**
- es un modelo **generativo**, lo que significa que se utiliza para **generar nuevas secuencias de datos** que se **asemejan** a los **datos de entrenamiento**

**ChatGPT** es una **herramienta de procesamiento de lenguaje natural**, de tipo **texto a texto**, que está **basada** en modelos **GPT**.

- Ha sido diseñado para generar texto en lenguaje natural con fines conversacionales (según OpenAI, ChatGPT es muy similar a su **modelo anterior, InstructGPT**; no obstante, InstructGPT está diseñado para generar texto instructivo para tareas como responder preguntas o proporcionar orientación paso a paso).

Algunas características de los modelos subyacentes es que se basan en la obtención del **siguiente token**, se han **dejado** de **entrenar** en **2021**, **no** poseen conexión a **internet**, etc.

Posible **GAI**

# ChatGPT | Limitaciones

Respuestas incorrectas o **sin sentido** como **totalmente verdaderas**. Problema desafiante por las siguientes razones:

- durante el entrenamiento RL (Reinforcement Learning, RL), actualmente, **no existe una fuente de verdad**;
- **entrenar** al modelo **para** que **sea precavido provoca** que **rechace** preguntas que puede responder correctamente;
- y el entrenamiento supervisado confunde al modelo porque la **respuesta ideal depende de lo que el modelo sabe**, en lugar de lo que sabe el humano

**Sensible a variaciones en el prompt** de la entrada o al intentar usar el mismo prompt varias veces

- Por ejemplo, dado un enfoque de una pregunta, el modelo puede afirmar que no sabe la respuesta, pero realizando cambios superficiales, puede responder correctamente

Modelo, a menudo, excesivamente **verboso** y **repetitivo** (por sesgos en entrenamiento y sobreoptimización)

# ChatGPT | Limitaciones

El modelo se centra en **averiguar lo que el usuario pretende decir** en lugar de hacer preguntas aclaratorias cuando el usuario proporciona una consulta ambigua

## Solicitudes inapropiadas:

- El modelo rechaza solicitudes inapropiadas; no obstante, indican que a veces responderá a instrucciones dañinas o mostrará un comportamiento sesgado
- Se utiliza la llamada **Moderation API** para advertir o bloquear ciertos tipos de contenido inseguro, no obstante, no es del todo precisa, por lo que esperan obtener algunos falsos negativos y positivos



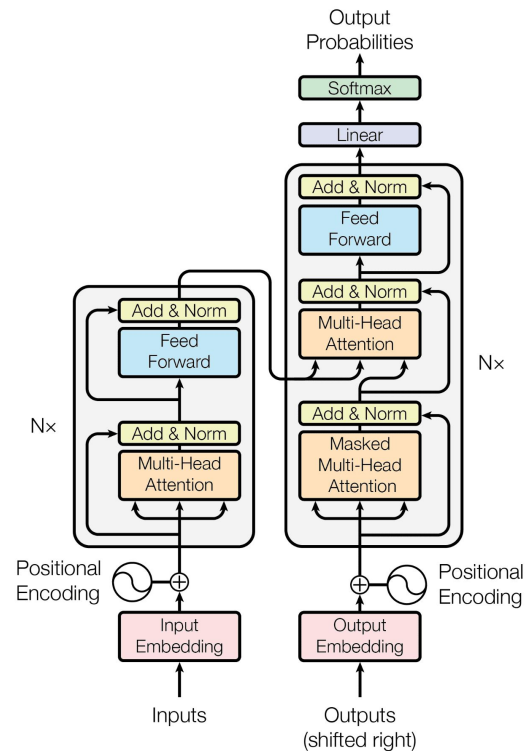


# ChatGPT | Arquitectura

En **2017** aparecen los **transformers**, un tipo de red neuronal artificial que **evoluciona** otros modelos de procesamiento de secuencias como la redes neuronales recurrentes (**RNN**) y la redes de memoria a largo corto plazo (**LSTM**)

A diferencia de los dos tipos de redes neuronales mencionadas, los **transformers no procesan** los datos de las secuencias **de forma secuencial**, sino que procesan toda la secuencia al mismo tiempo prestando "atención" a las palabras más importantes. Este mecanismo de "atención" les permite concentrarse más en las palabras relevantes y menos en las palabras menos importantes

Además, los transformers **pueden procesar toda la secuencia al mismo tiempo**



# ChatGPT | Arquitectura

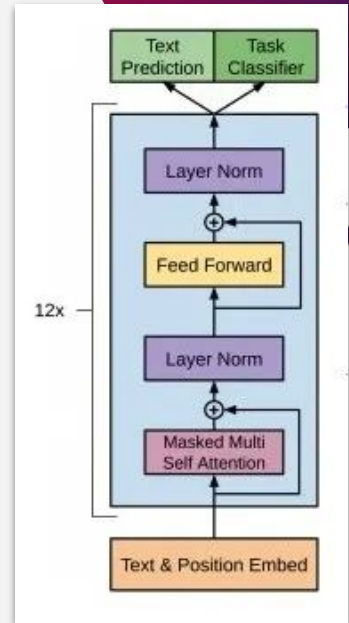
**GPT-3** se ha entrenado con una cantidad ingente de datos, aproximada a los **45 TB**.

Al momento de su lanzamiento, era la red neuronal artificial más grande existente de cara al público con **175.000 millones** de parámetros.

Para entrenar el modelo:

- Utilizamos un conjunto de datos con **millones de tokens** para generar ejemplos de entrenamiento para el modelo.
- Mostramos el **vector de características** (frase a la que le vamos a añadir la palabra) y enseñamos la **palabra correcta** para la salida (etiqueta)
  - Si la **palabra de salida** es **diferente** a la que esperamos **calculamos** el **error** mostrando la salida correcta y **actualizamos** los **parámetros** del modelo, para que la próxima vez haga una predicción mejor.

Repetimos este proceso multitud de veces



Text: Second Law of Robotics: A robot must obey the orders given it by human beings

Generated training examples

Example #	Input (features)	Correct output (labels)
1	Second law of robotics :	a
2	Second law of robotics : a	robot
3	Second law of robotics : a robot	must
...		

# ChatGPT | Técnicas

Reinforcement Learning  
from Human Feedback,  
**RLHF**

## Modelo Supervised Fine-Tuning (SFT)

Consiste en recopilar datos de demostración para **entrenar** un modelo de política supervisada, denominado **modelo SFT**

- Se selecciona una **lista de prompts** y se pide a un grupo de etiquetadores **humanos** que **escriban** la **respuesta** de salida esperada

En el caso de ChatGPT, se han utilizado dos fuentes distintas de instrucciones:

- algunas han sido preparadas **directamente** por los **etiquetadores** o desarrolladores,
- y otras se han extraído de las peticiones de la **API** de OpenAI (es decir, de sus clientes de GPT-3)

**Produce** → Conjunto de datos relativamente pequeño y de alta calidad

**Problema** → Alto coste / Poca escalabilidad

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 learning.

↓  
SFT  
🧠  
📄📄📄

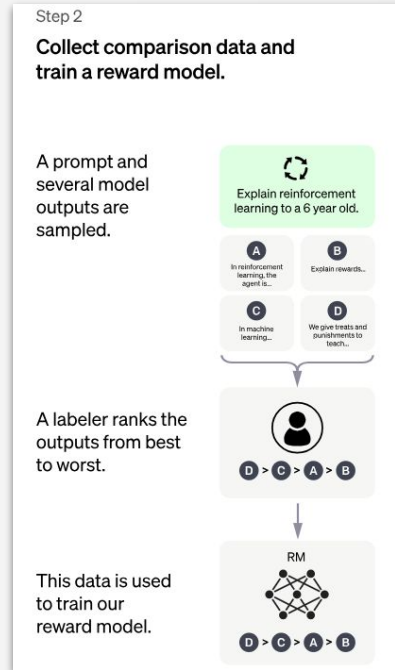
# ChatGPT | Técnicas

## Modelo de recompensa (RM)

Consiste en **ordenar varias respuestas para el mismo prompt** en relación a las que sean preferibles para los humanos:

- Se selecciona una **lista de preguntas** y el modelo SFT genera varios **resultados** (entre **4 y 9**) para cada **pregunta**.
- Los **etiquetadores clasifican** los resultados de **mejor a peor**. El resultado es un **nuevo conjunto de datos etiquetados**, en el que las clasificaciones son las etiquetas. El tamaño de este conjunto de datos es aproximadamente **10 veces mayor** que el conjunto de datos curados utilizado para el modelo SFT.
- Estos nuevos datos se utilizan para **entrenar un modelo de recompensa** (Reward Model, RM). Este modelo toma como **entrada algunos de los resultados del modelo SFT** y los **clasifica por orden de preferencia**.

**Objetivo** → Extraer de los datos un sistema automático que se supone que imita las preferencias humanas (función objetivo, RM)



# ChatGPT | Técnicas

## Fine-tuning del modelo SFT usando Proximal Policy Optimization (PPO)

Consiste en afinar la política de SFT para obtener un modelo PPO usando el modelo de recompensa.

El algoritmo específico utilizado se denomina **Proximal Policy Optimization (PPO)** y el modelo ajustado se denomina modelo PPO. PPO es:

- Un **algoritmo** que se utiliza para entrenar agentes en el aprendizaje por refuerzo ("**on-policy**", adapta continuamente la política actual).
- Utiliza un método de **optimización de región** de confianza para entrenar la política.
- Utiliza una **función de valor** para estimar el rendimiento esperado de un estado o acción

Dado el **prompt** y la **respuesta**, se produce una **recompensa** (determinada por el *modelo de recompensa*). Se añade una penalización KL por token del modelo SFT en cada token para mitigarla sobreoptimización del modelo de recompensa.

**Produce** → Modelo PPO

Step 3

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

A new prompt is sampled from the dataset.



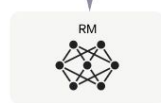
The PPO model is initialized from the supervised policy.



The policy generates an output.

Once upon a time...

The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.





# 03

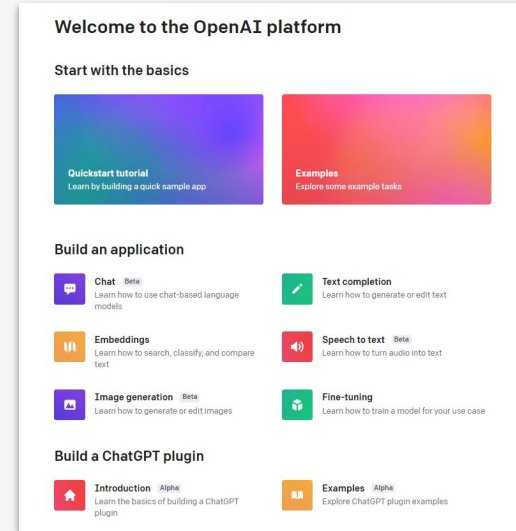
## Aplicaciones

# Aplicaciones

**Multitud** de **usos** posibles (aproximación a **GAI**)

Modelos expuestos a través de una **API**

**Plataforma** con cursos, guías documentación y ejemplos reales sobre diversos temas



# Aplicaciones

## Asistentes de código

### Prompt

```
### Postgres SQL tables, with their properties:
#
# Employee(id, name, department_id)
# Department(id, name, address)
# Salary_Payments(id, employee_id, amount, date)
#
### A query to list the names of the departments which employed
more than 10 employees in the last 3 months
SELECT
```

### Sample response

```
SELECT d.name
FROM Department d
INNER JOIN Employee e ON d.id = e.department_id
INNER JOIN Salary_Payments sp ON e.id = sp.employee_id
WHERE sp.date > NOW() - INTERVAL '3 months'
GROUP BY d.name
HAVING COUNT(*) > 10
```

## Generador de resúmenes

### Prompt

Convert my short hand into a first-hand account of the meeting:

Tom: Profits up 50%  
Jane: New servers are online  
Kjel: Need more time to fix software  
Jane: Happy to help  
Parkman: Beta testing almost done

### Sample response

At the meeting, Tom reported that profits had increased by 50%. Jane then mentioned that the new servers were online. Kjel mentioned that they needed more time to fix the software, and Jane offered to help. Finally, Parkman reported that the beta testing was almost done.

## Recetas a partir de ingredientes

### Prompt

Write a recipe based on these ingredients and instructions:

Frito Pie

Ingredients:

Fritos  
Chili  
Shredded cheddar cheese  
Sweet white or red onions, diced small  
Sour cream

Instructions:

### Sample response

1. Preheat oven to 350 degrees F.
2. Spread a layer of Fritos in the bottom of a 9x13 inch baking dish.
3. Top the Fritos with a layer of chili.
4. Sprinkle the chili with shredded cheddar cheese.
5. Sprinkle the cheese with diced onions.
6. Bake in preheated oven for 25 minutes.
7. Remove from oven and top with a dollop of sour cream.
8. Serve warm. Enjoy!





04

Prompt  
engineering

# Prompt engineering

Si preguntamos a ChatGPT **¿qué es la prompt engineering?**, obtenemos la siguiente respuesta:

*El término prompt engineering refiere a la ciencia y **arte** de **diseñar, estructurar** y **optimizar prompts** (indicaciones o estímulos) **para obtener las respuestas más adecuadas de un modelo de lenguaje generativo**, como GPT-4 de OpenAI. **[ChatGPT]***

**No sólo** se limita a la construcción de la **instrucción inicial**, sino que también abarca el **manejo de las respuestas del modelo** y la **preparación de prompts de seguimiento**, para mantener un flujo de conversación natural y consistente → Importante en la **creación de aplicaciones**

Surge un nuevo perfil profesional, **prompt engineer** (desde el punto de vista del consumidor del modelo); y formaciones y **cursos**: entre ellos, el impartido por los reconocidos **Andrew Ng** (DeepLearning.AI y Coursera) e **Isa Fulford** (OpenAI), **ChatGPT Prompt Engineering for Developers**

# 05

## Plugins



# Plugins

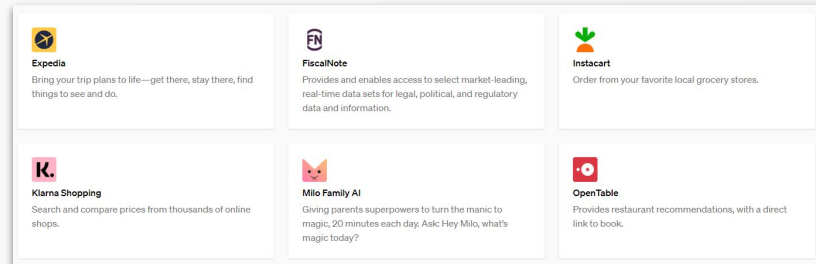
¡Nos dan la posibilidad de **eliminar limitaciones** y **expandir la funcionalidad** con **conexión a internet, procesamiento de documentos y vídeos, etc!**

## Algunos plugins:

**ChatWithPDF:** nos permite **interactuar con archivos PDF** como si estuviéramos hablando con un humano. Pegamos el enlace URL a cualquier archivo PDF y hacemos preguntas sobre su contenido.

**Video Insights:** nos da la posibilidad de pegar la **URL de cualquier vídeo de YouTube** dentro de nuestra pregunta. Podemos pedir un breve resumen del vídeo o hacer cualquier pregunta sobre el contenido del mismo.

**Kayak:** nos da la posibilidad de hacer **preguntas sobre vuelos, hoteles, alquiler de coches y actividades** con el objetivo de obtener recomendaciones personalizadas basadas en nuestras preferencias y en los datos disponibles en el conocido buscador Kayak. Puede ayudarnos a encontrar las mejores ofertas y a reservar nuestro viaje de forma rápida y cómoda.

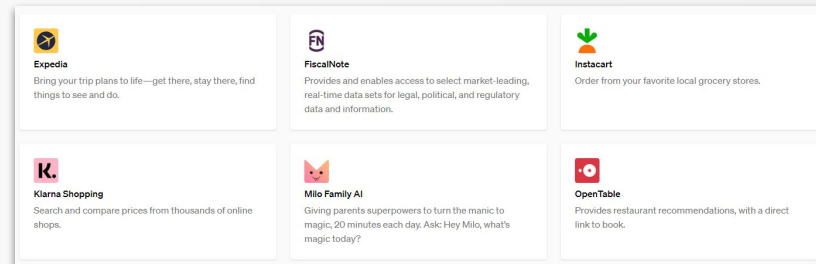


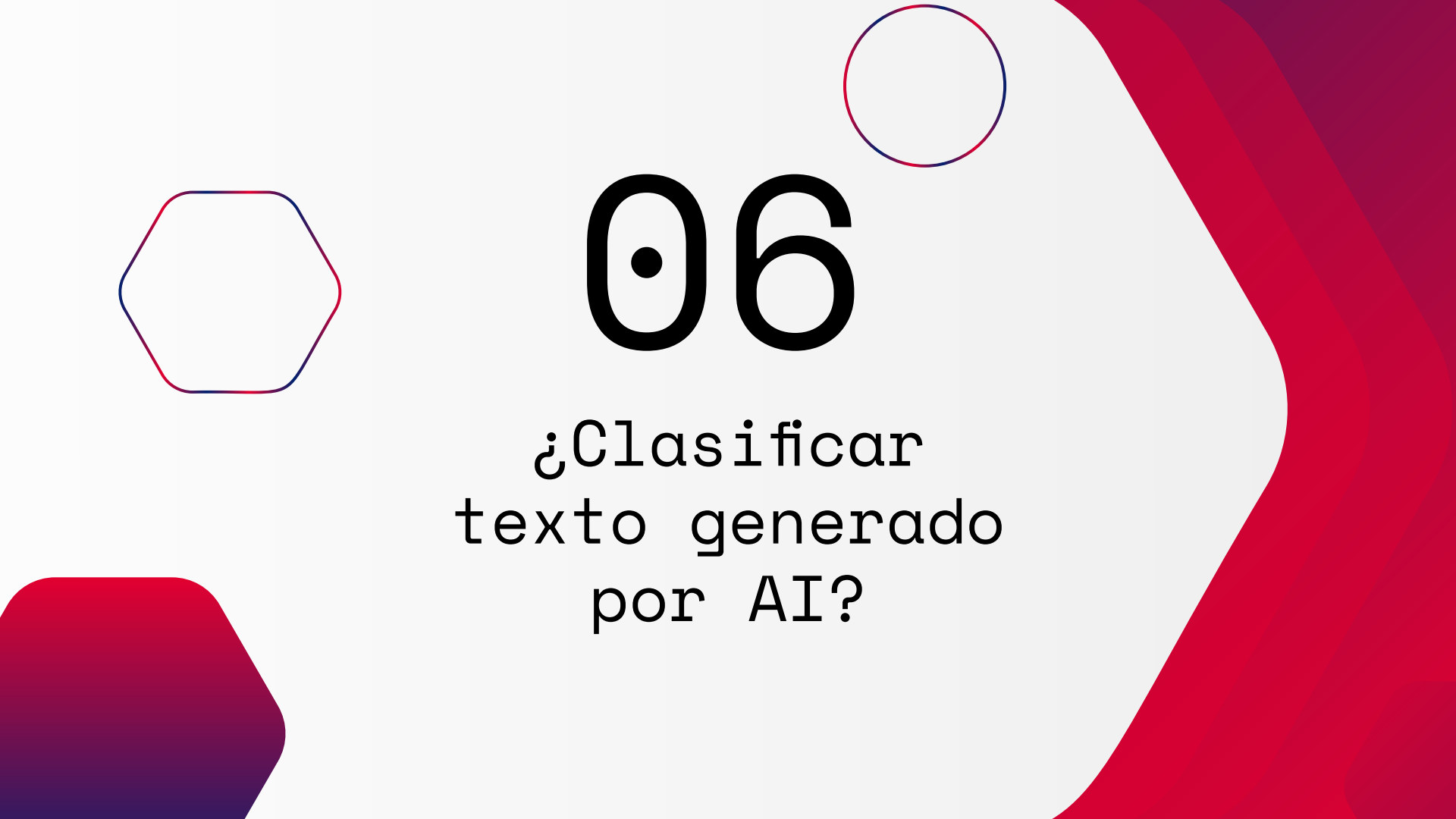
# Plugins

**KeyMate AI Search:** nos ofrece un **modo de búsqueda sencillo y efectivo a través de Google**. Utiliza la API de búsqueda de Google para escanear y resumir rápidamente los principales resultados de búsqueda para una palabra clave determinada. Básicamente, analiza resúmenes de texto de todos los resultados de la primera página de búsqueda y proporciona una respuesta actualizada basada en los mismos.

**Lexi Shopper:** nos ofrece **recomendaciones de productos de Amazon**. Podemos pedir a ChatGPT que nos muestre productos en función de nuestras preferencias, presupuesto o necesidades. Utiliza los datos de Amazon para acceder a millones de productos y ofrecernos enlaces y precios.

**OpenTable:** nos facilita la **búsqueda y reserva de restaurantes a nuestra petición**. Podemos pedirle que nos muestre restaurantes en función de nuestra ubicación, cocina, ocasión, presupuesto u otras preferencias. A partir de la base de datos de OpenTable, que cuenta con más de 60.000 restaurantes en todo el mundo, nos ofrece recomendaciones y enlaces para reservar.





# 06

¿Clasificar  
texto generado  
por AI?

# ¿Clasificar texto generado por AI?

OpenAI, con el objetivo de desmentir las falsas afirmaciones de que el texto generado con IA ha sido escrito por un humano, por ejemplo:

- en la realización de campañas automatizadas de desinformación,
- en el uso de herramientas de IA para trabajos académicos
- o el posicionamiento de un chatbot de IA como si fuera un humano;

ha creado una herramienta denominada **AI Text Classifier**

## AI Text Classifier

The AI Text Classifier is a fine-tuned GPT model that predicts how likely it is that a piece of text was generated by AI from a variety of sources, such as ChatGPT.

This classifier is available as a free tool to spark discussions on AI literacy. For more information on ChatGPT's capabilities, limitations, and considerations in educational settings, please visit [our documentation](#).


### Current limitations:

- Requires a minimum of 1,000 characters, which is approximately 150 - 250 words.
- The classifier isn't always accurate; it can mislabel both AI-generated and human-written text.
- AI-generated text can be edited easily to evade the classifier.
- The classifier is likely to get things wrong on text written by children and on text not in English, because it was primarily trained on English content written by adults.


### Try the classifier

To get started, choose an example below or paste the text you'd like to check. Be sure you have appropriate rights to the text you're pasting.

### Examples

 Human-Written

 AI-Generated

 Misclassified Human-Written

### Text

Enter your document text here

# ¿Clasificar texto generado por IA?

Este clasificador nos indica **si existe posibilidad de que un texto se haya generado usando un modelo de AI**. Las **clases** en las que clasifica son las siguientes:

*Very unlikely to be AI-generated,*

*Unlikely to be AI-generated,*

*Unclear if it is AI written,*

*Possibly AI-generated,*

*Likely AI-generated*

**Limitaciones** actuales:

- Requiere un **mínimo de 1.000 caracteres**
- El clasificador **no** siempre es **preciso**
- El **texto** generado por IA puede ser **editado** fácilmente para **evadir al clasificador**
- **Fue entrenado** con contenido en **inglés** escrito por adultos





07

Conclusiones

# Conclusiones

- ChatGPT **mejora múltiples aspectos** de nuestra vida cotidiana (organización de un viaje, aprendizaje, copywriting, interacción, etc.)
- **Hito** en la historia del desarrollo de la **inteligencia artificial** (gran visibilización)
- Herramienta global muy extendida
- **Potencial** innegable y nueva forma de construir
- Implicaciones **éticas** y de **seguridad**
  - ◆ Riesgo por **mal uso** (generación de **desinformación** o contenido **dañino**)
  - ◆ **Privacidad** y potencial **sesgo** en los **datos** de entrenamiento
  - ◆ **Políticas** adecuadas para **prevenir** y **eliminar** estos **riesgos**
- **Futuro** basado en **reducción de costes**, expansión del modelo, adopción de mejoras, **mayor accesibilidad**



08

Bibliografía

# Bibliografía

- **[OpenAI (2023) Introducing ChatGPT]**  
<https://openai.com/blog/chatgpt>
- **[Some Glimpse AGI in ChatGPT. Others Call It a Mirage]**  
<https://www.wired.com/story/chatgpt-agi-intelligence/>
- **[AI vs. Machine Learning vs. Deep Learning vs. Neural Networks: What's the Difference?]**  
<https://www.ibm.com/cloud/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks>
- **[Apuntes SIGE]**  
Apuntes de la asignatura de Sistemas Inteligentes para la Gestión en la Empresa del Máster Profesional en Ingeniería Informática de la Universidad de Granada
- **[Attention Is All You Need]**  
<https://dl.acm.org/doi/pdf/10.5555/3295222.3295349>
- **[GPT-3]**  
<https://en.wikipedia.org/wiki/GPT-3>
- **[The Transformer Model]**  
<https://machinelearningmastery.com/the-transformer-model/>



# Bibliografía

- **[Chat GPT and GPT 3 Detailed Architecture Study-Deep NLP Horse]**  
<https://medium.com/nerd-for-tech/gpt3-and-chat-gpt-detailed-architecture-study-deep-nlp-horse-db3af9de8a5d>
- **[How Does ChatGPT Actually Work?]**  
<https://www.scalablepath.com/data-science/chatgpt-architecture-explained>
- **[How ChatGPT actually works]**  
<https://www.assemblyai.com/blog/how-chatgpt-actually-works/>
- **[How ChatGPT Works Technically | ChatGPT Architecture]**  
<https://www.youtube.com/watch?v=bSvTVREwSNw>
- **[OpenAI Models]**  
<https://platform.openai.com/docs/models>
- **[Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models]**  
<https://arxiv.org/pdf/2304.01852.pdf>
- **[ChatGPT: Everything you need to know about OpenAI's GPT-4 tool]**  
<https://www.sciencefocus.com/future-technology/gpt-3/>



# Bibliografía

- **[How Transformers Work]**  
<https://towardsdatascience.com/transformers-141e32e69591>
- **[New and improved content moderation tooling]**  
<https://openai.com/blog/new-and-improved-content-moderation-tooling>
- **[Pricing]**  
<https://openai.com/pricing>
- **[OpenAI platform]**  
<https://platform.openai.com/>
- **[Platform OpenAI Examples]**  
<https://platform.openai.com/examples>
- **[ChatGPT]**  
<https://chat.openai.com/>
- **[ChatGPT plugins]**  
<https://openai.com/blog/chatgpt-plugins>



# Bibliografía

- **[ChatGPT Prompt Engineering for Developers]**  
<https://www.deeplearning.ai/short-courses/chatgpt-prompt-engineering-for-developers/>
- **[The 12 Best ChatGPT Plugins]**  
<https://approachableai.com/best-chatgpt-plugins/>
- **[AI Text Classifier]**  
<https://platform.openai.com/ai-text-classifier>



# ¡ Gracias !

¿Alguna pregunta?



**CREDITS:** This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**

Please keep this slide for attribution