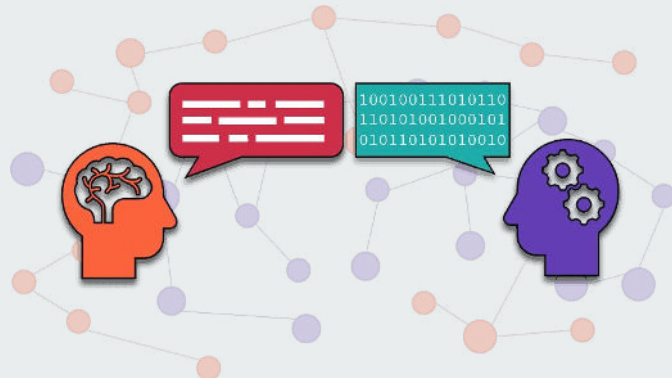


Maite Martín (maite@ujaen.es)

Combatiendo el Discurso de Odio mediante la Generación Automática de Contranarrativas usando IA



Ciclo de Charlas ¿ De qué estamos hablando?
3 de marzo de 2025



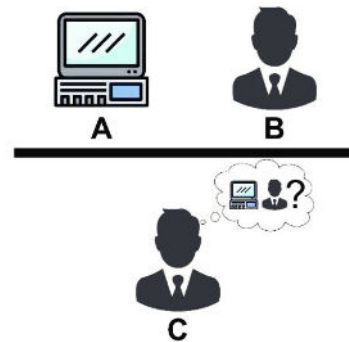


**¿Qué es la
Inteligencia
Artificial?**

Sistemas computacionales que imitan el comportamiento humano



**Test de
Turing**
(1950)



Conferencia de Dartmouth

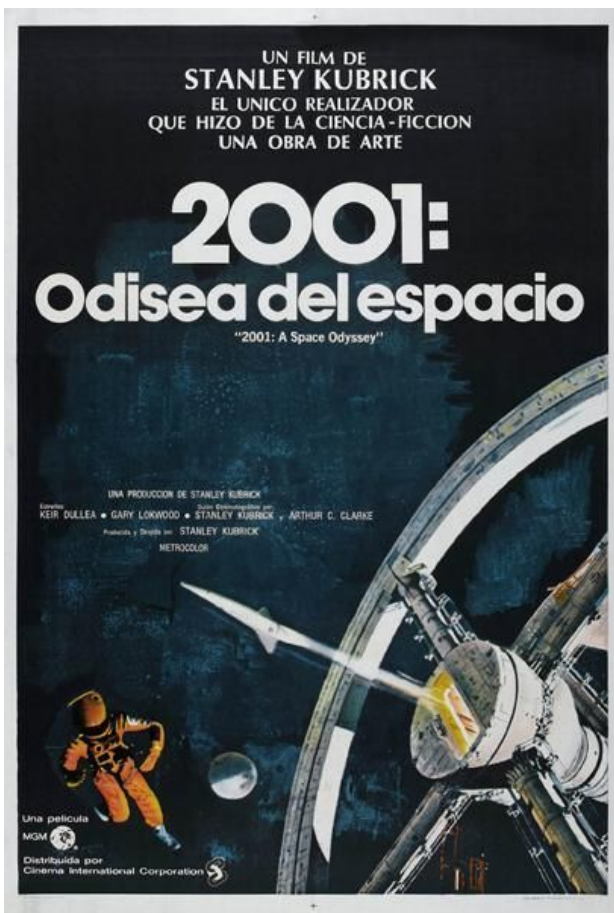
En el verano de 1956, un grupo de científicos, liderados por John McCarthy, Marvin Minsky, Nathaniel Rochester y Claude Shannon, se reunió en Dartmouth College (New Hampshire) para discutir ideas que marcarían el inicio de una nueva disciplina que denominaron **“inteligencia artificial”** y que se fundamentaba en que **“cada aspecto del aprendizaje o cualquier otra característica de la inteligencia humana puede, en principio, ser descrita de manera tan precisa que una máquina podría simularla”**



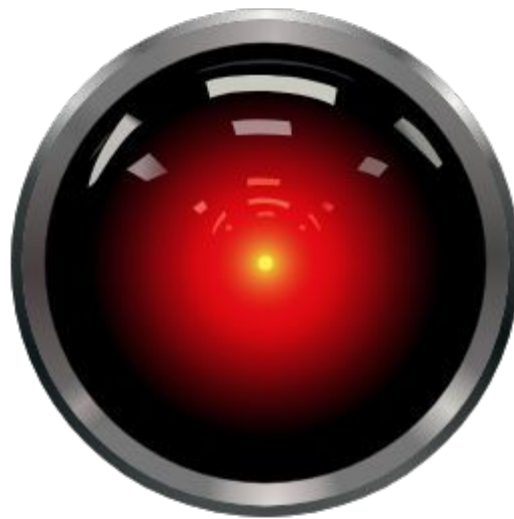


Procesamiento del lenguaje natural

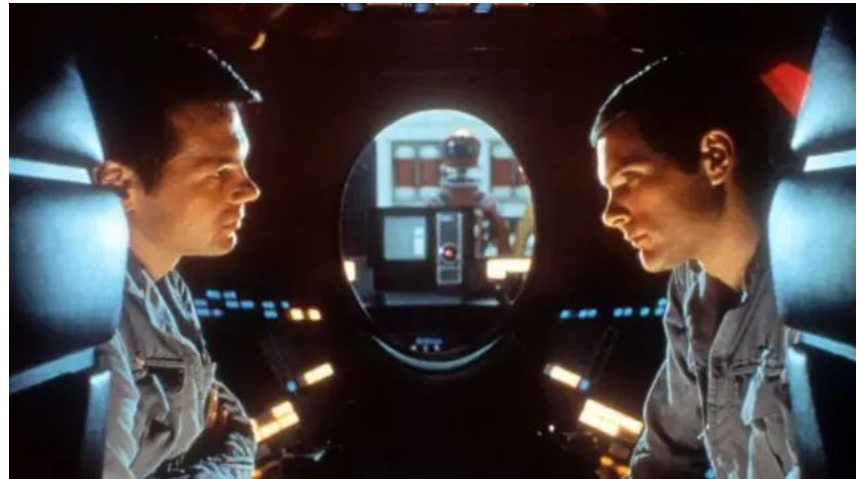


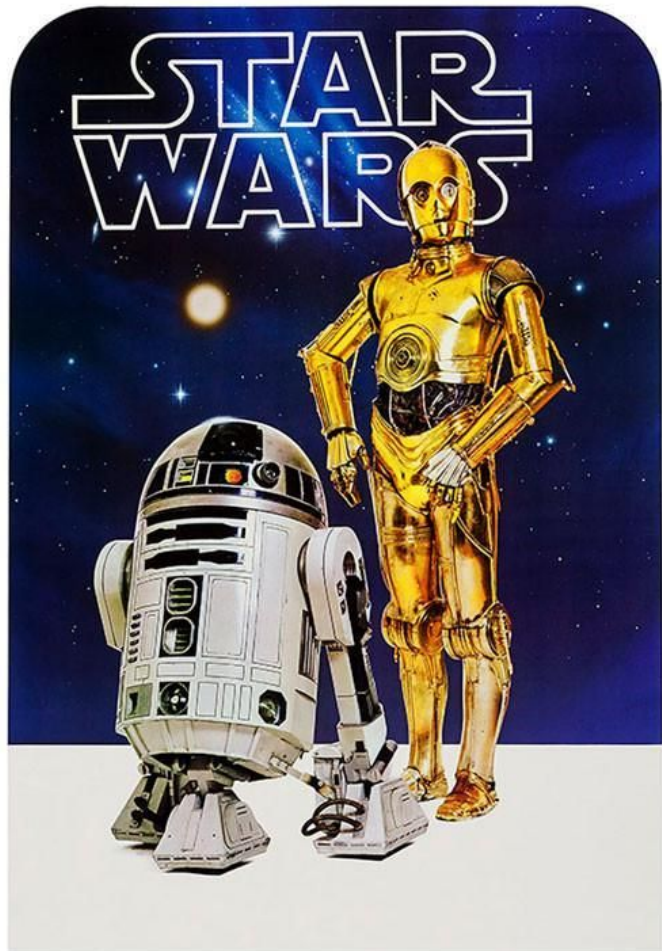


1968



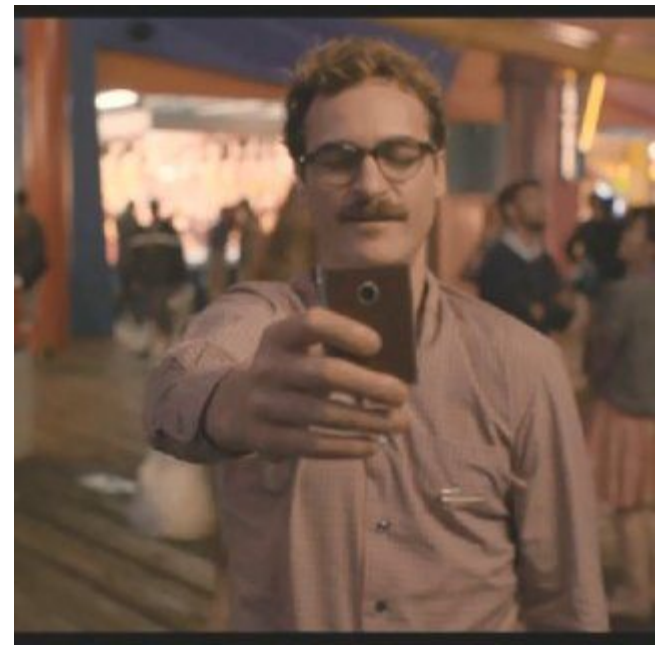
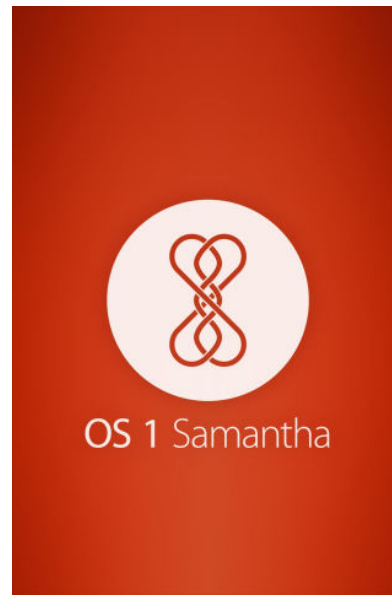
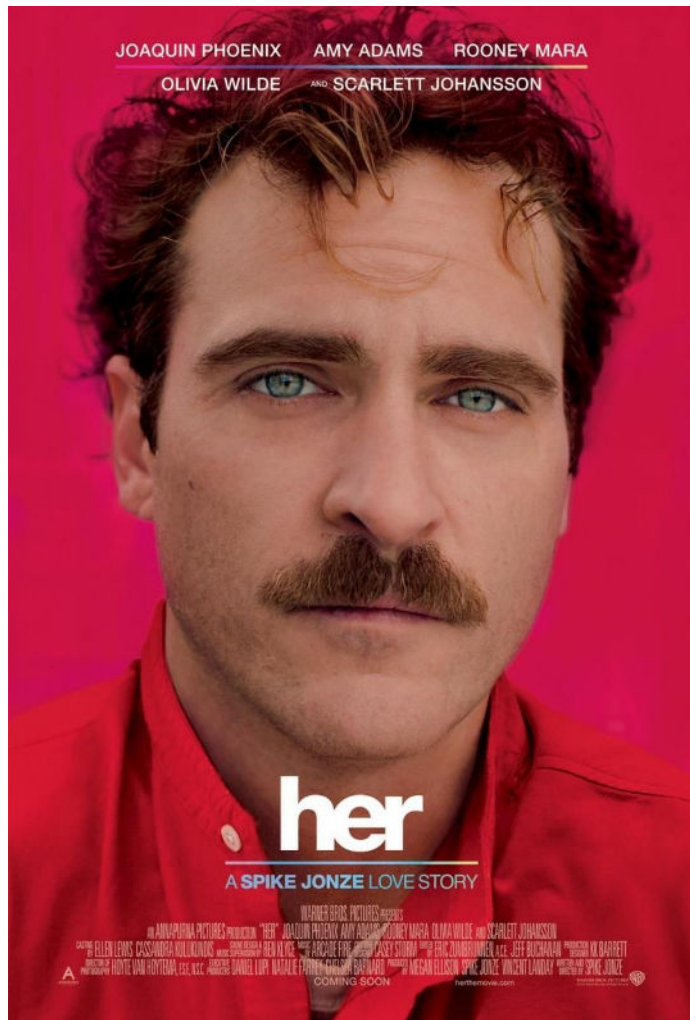
'HAL 9000':
*Heuristically programmed
ALgorithmic computer o
una referencia al gigante
de la computación de
aquella época (IBM)*





C-3PO es un personaje icónico de la saga de *Star Wars*. Es un droide de protocolo diseñado para asistir en tareas de traducción e interpretación, ya que domina más de seis millones de formas de comunicación.

1977



2013

Inteligencia Artificial

**Procesamiento del
Lenguaje Natural (PLN)**

**Grandes Modelos del
Lenguaje (LLM)**



**ChatGPT es un sistema de diálogo basado en PLN
que gestiona conversaciones de manera contextual y
coherente**

[illegible]

LC: Lingüística Computacional
CL: Compresión del Lenguaje

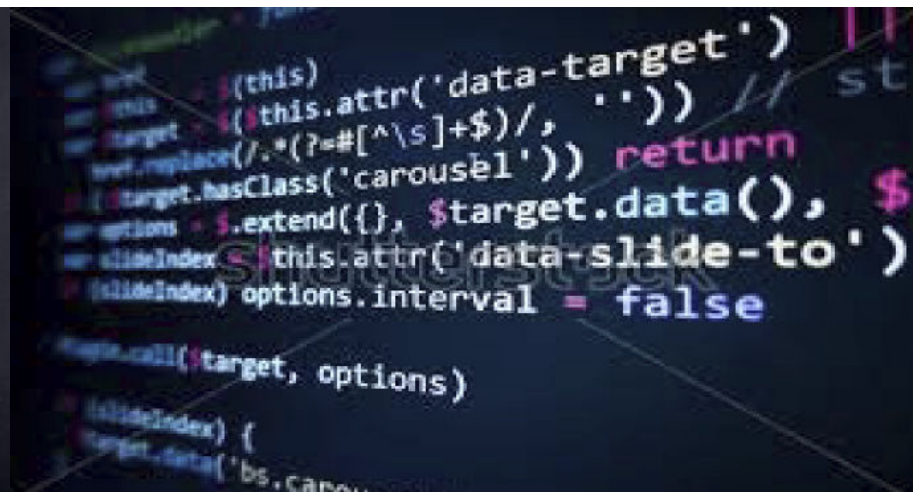
CL: Computational Linguistic
NLU: Natural Language Understanding

HLT: Human Language Technologies

NLP: Natural Language Processing

Lenguaje natural y artificiales

- Lenguajes naturales: español, gallego, inglés, alemán, francés, sumerio, latín, griego antiguo, árabe...
- Lenguajes artificiales: lenguajes de programación, escala musical, código Morse, esperanto





55%
De la IA es PLN

¿Cuál es el objetivo final?



Que los ordenadores generen y entiendan el lenguaje igual que lo hacen los humanos

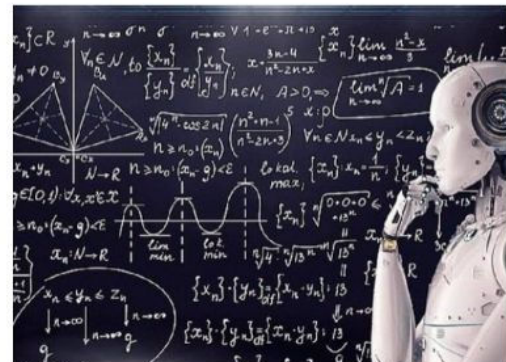
¿Cuál es el objetivo final?



Que los ordenadores **generen** y **entiendan** el lenguaje igual que lo hacen los humanos

ChatGPT aprueba el MIR 2023 con solo un 25% de respuestas incorrectas

Tutores de CTO analizan el examen de esta inteligencia artificial con un muestrario de 20 preguntas



El chat se ha sometido a un examen MIR de 20 preguntas

Por qué los programadores están en realidad encantados con el auge de ChatGPT

Emilio David, Business Insider 5 mar. 2023 7:00h.



APRENDIZAJE AUTOMÁTICO

ChatGPT revoluciona las aulas de arriba a abajo



- La llegada de un modelo de Inteligencia Artificial que redacta como un humano y resuelve problemas está impactando ya en los colegios y las universidades
- Las universidades ven futuras ventajas, pero no saben cómo van a abordar el problema del plagio a partir de ahora
- Así usan los estudiantes ChatGPT: ejemplos de buenas prácticas



TECNOLOGÍAS DISRUPTIVAS

La Inteligencia Artificial revoluciona la ciencia y la docencia

ChatGPT ya publica artículos científicos y es mejor que un MBA medio: lo cambia todo

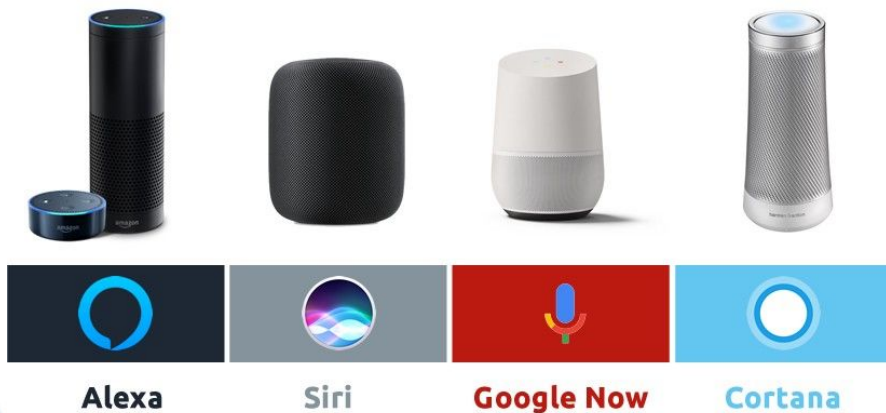
Agentes conversacionales (ChatBot)



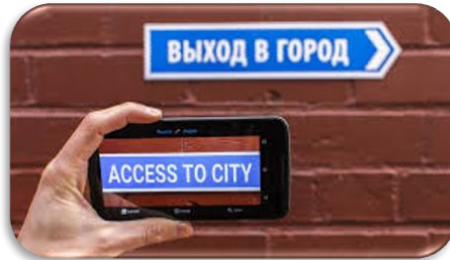
Ejemplos de uso diario

—

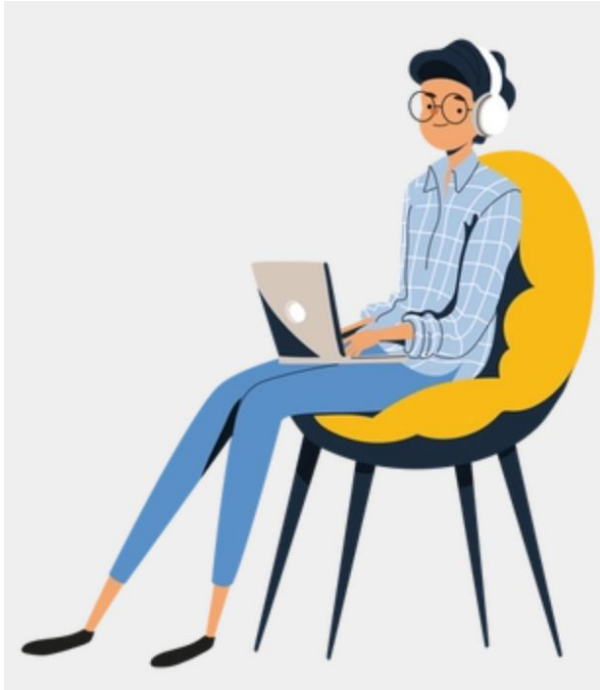
Asistentes de voz



Traductores automáticos



Predictores de texto



New Message

Ann Doe (ann_doe@gmail.com)

Work Agenda

Hello! I wanted to ask about our la|st meeting



Untitled document



File Edit View Insert Format Tools Add-ons Help Last edit was seconds ago



100%

Normal text

Arial



11



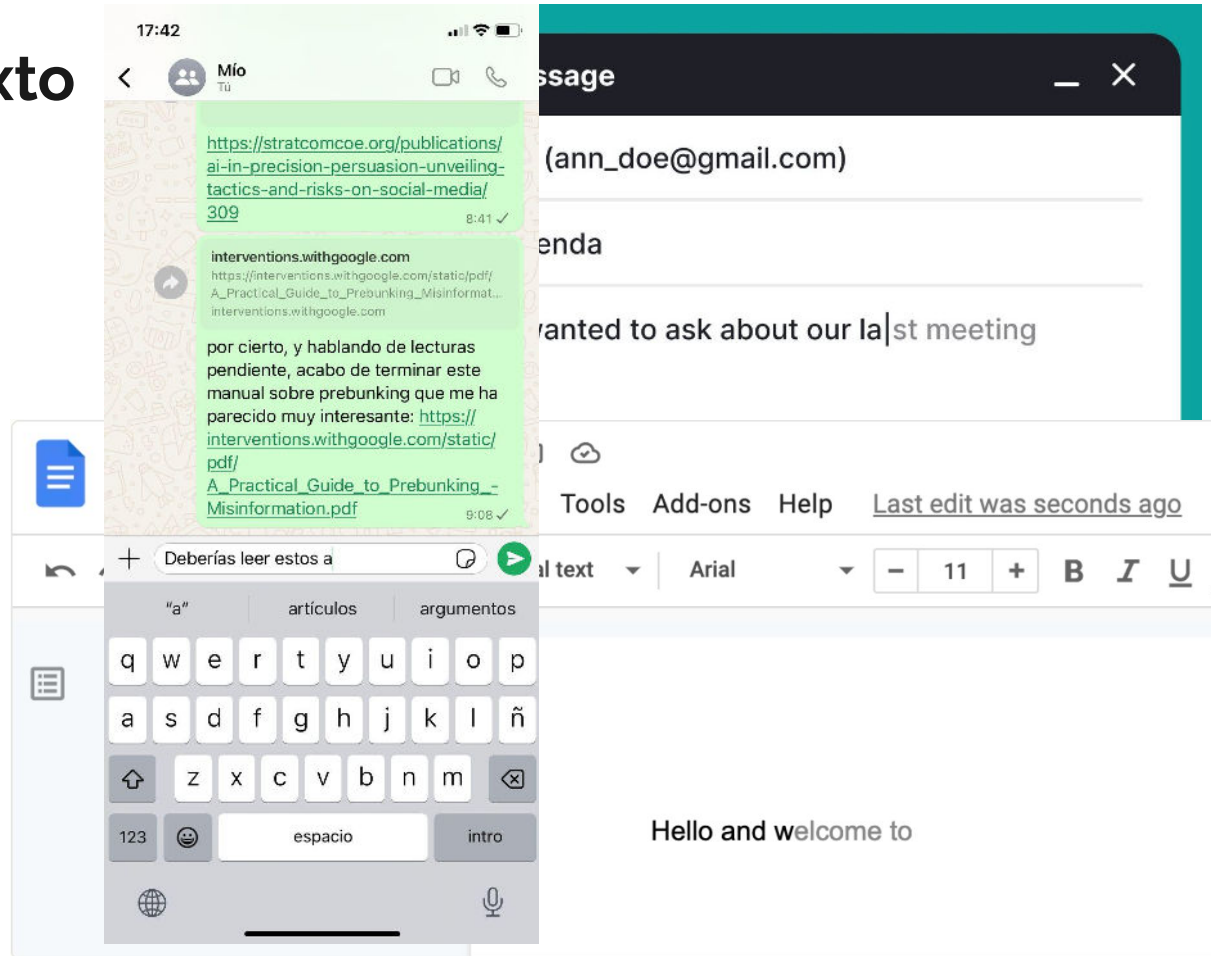
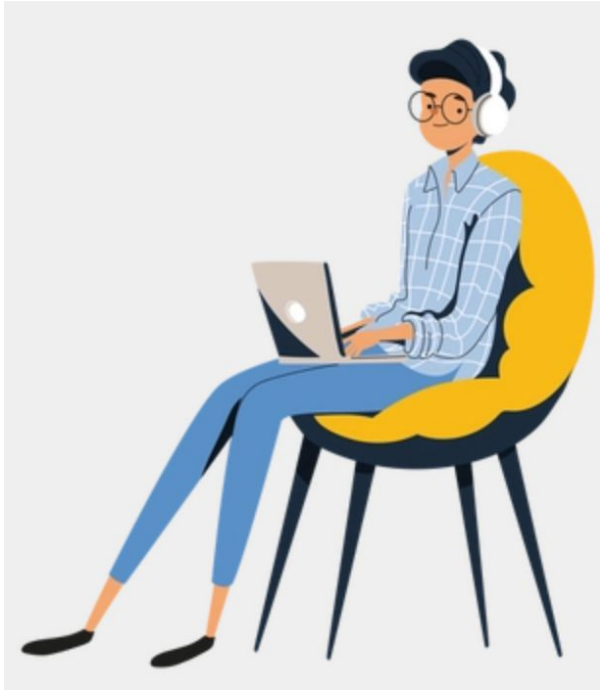
B

I

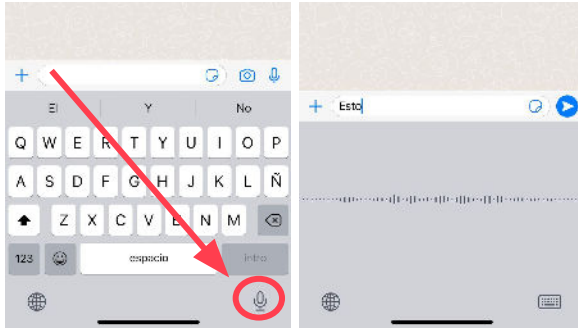
U

Hello and welcome to

Predictores de texto



Transcripción y corrección



Google





Cualquier aplicación
que requiera el uso
masivo de texto
desestructurado





20%

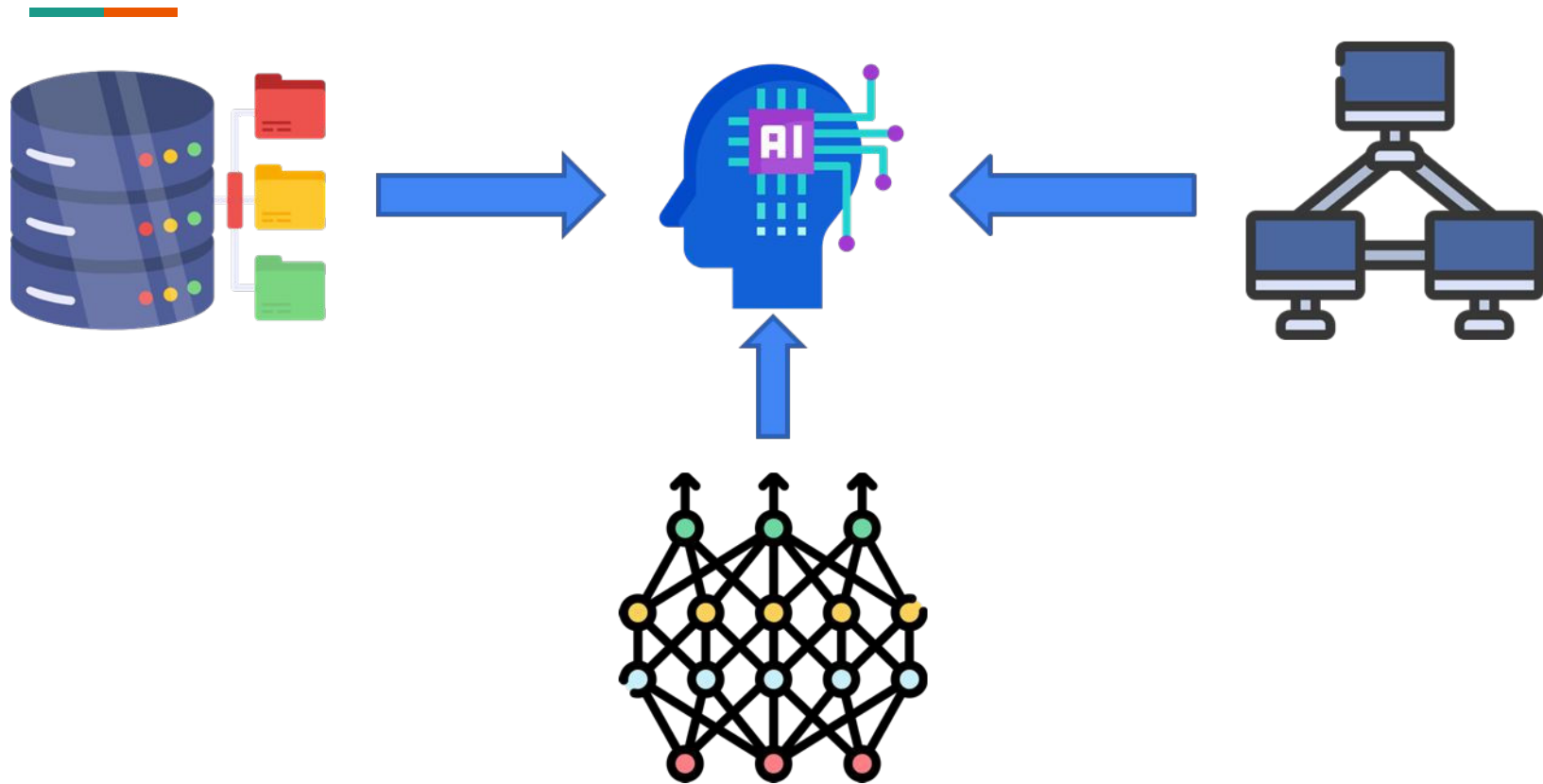
Información estructurada

80%

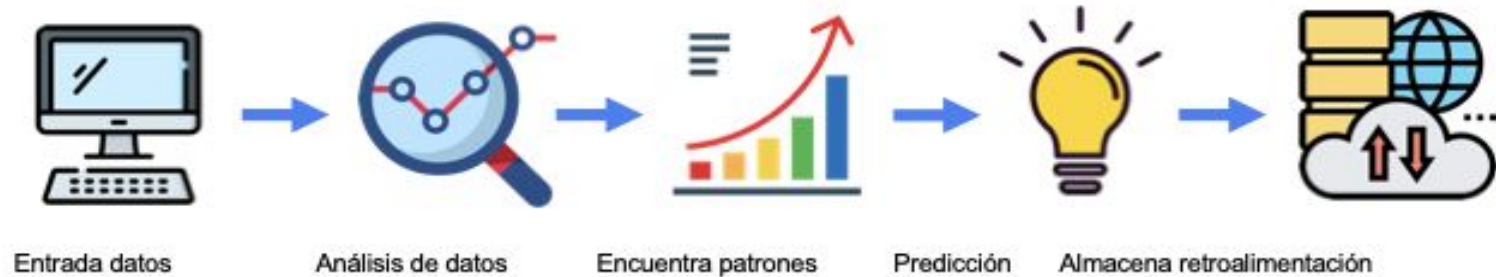
**Información
desestructurada**

¿Por qué ahora?

¿Por qué ahora?



Aprendizaje automático (Machine Learning)



Un algoritmo de ML se entrena en lugar de programarse

Se presentan múltiples ejemplos para la tarea que se desea aprender y el algoritmo encuentra una estructura estadística en estos ejemplos que eventualmente genera unas reglas que permiten automatizar la tarea

La IA y los modelos del lenguaje están de moda

Premio nobel de física 2024

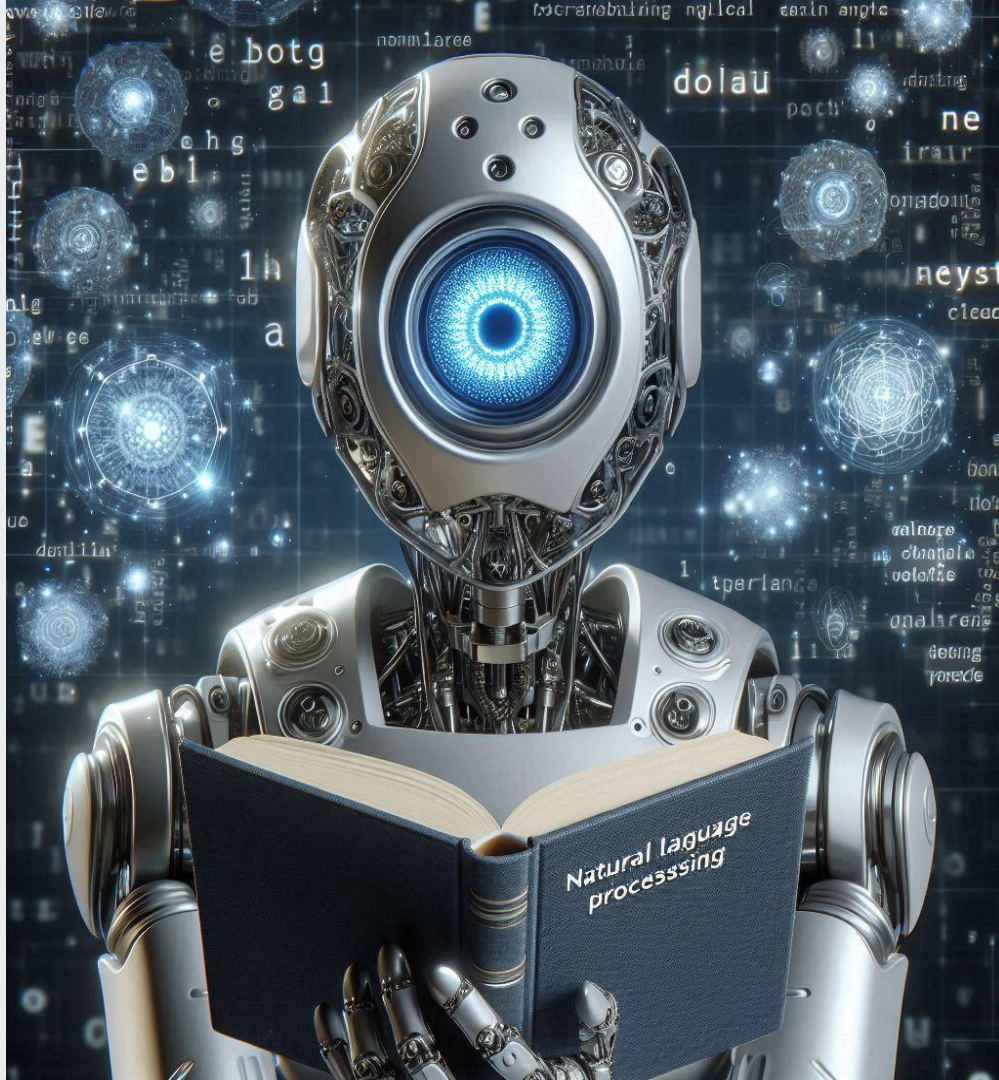


Premio nobel de química 2024



El Nobel de Química para el diseño computacional de proteínas y la predicción de su estructura

9 octubre 2024

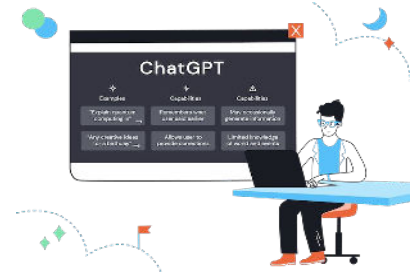


Qué es un LLM



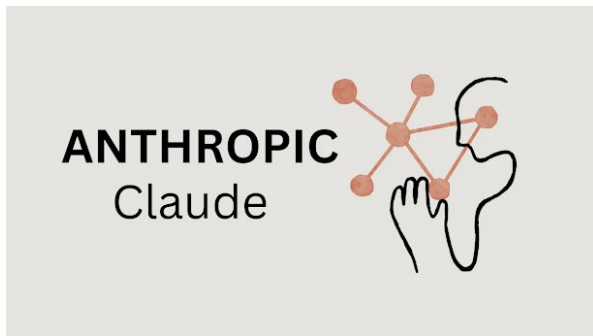
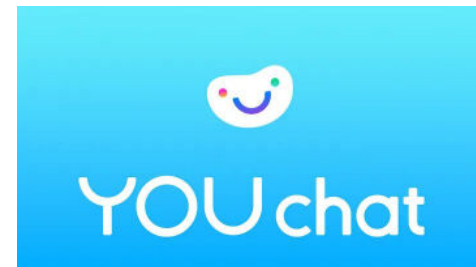
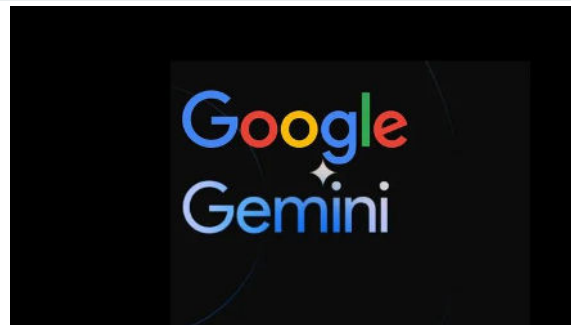
- Un LLM (Large Language Model) es un gran modelo del lenguaje que ha **aprendido a utilizar el lenguaje** como lo hacemos los humanos ya que ha sido **entrenado con enormes cantidades de texto**
- Los LLM funcionan al **aprender patrones en el texto que leen**. Por ejemplo, aprenden cómo se estructuran las oraciones, cómo se utilizan las palabras en diferentes contextos y cómo se conectan las ideas. Luego, pueden usar esta información para generar texto nuevo que tenga sentido

ChatGPT de OpenAI



A partir de una **gran cantidad de textos** de Internet (Wikipedia, webs, RRSS, Bibliotecas...), se aplican técnicas de redes neuronales profundas (**Generative Pretrained Transformer**) y ChatGPT ha aprendido a **predecir qué palabra es más probable** que siga a una secuencia dada

Otros LLMs



perplexity

The graphic features a stylized brain composed of a network of interconnected nodes and lines, with a central node containing the letter 'ñ'. To the right of the brain, the text 'PERTE' is written in large, bold, white capital letters, followed by 'Nueva Economía de la Lengua' in a smaller, white, sans-serif font. In the top right corner, there is a logo for the 'Plan de Recuperación, Transformación y Resiliencia' (Plan de Recuperación, Transformación y Resiliencia) with the text 'Plan de Recuperación, Transformación y Resiliencia' below it. In the bottom left corner, there are logos for the 'Ministerio de Educación y Formación Superior' and the 'Ministerio de Ciencia e Innovación'. In the bottom right corner, there is a logo for the 'Ministerio de Educación y Formación Superior' and the 'Ministerio de Ciencia e Innovación'.

ILENIA 
IMPULSO DE LAS LENGUAS
EN LA INTELIGENCIA ARTIFICIAL



El modelo fundacional abierto de IA generativa desarrollado en España

Red de centros ILENIA



**Barcelona
Supercomputing
Center**

Centro Nacional de Supercomputación



HiTZ

Hizkuntza Teknologia Zentroa
Basque Center for Language Technology



UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA



CENID

CENTRO DE
INTELIGENCIA DIGITAL
PROVINCIA DE ALICANTE



Universidad de Jaén

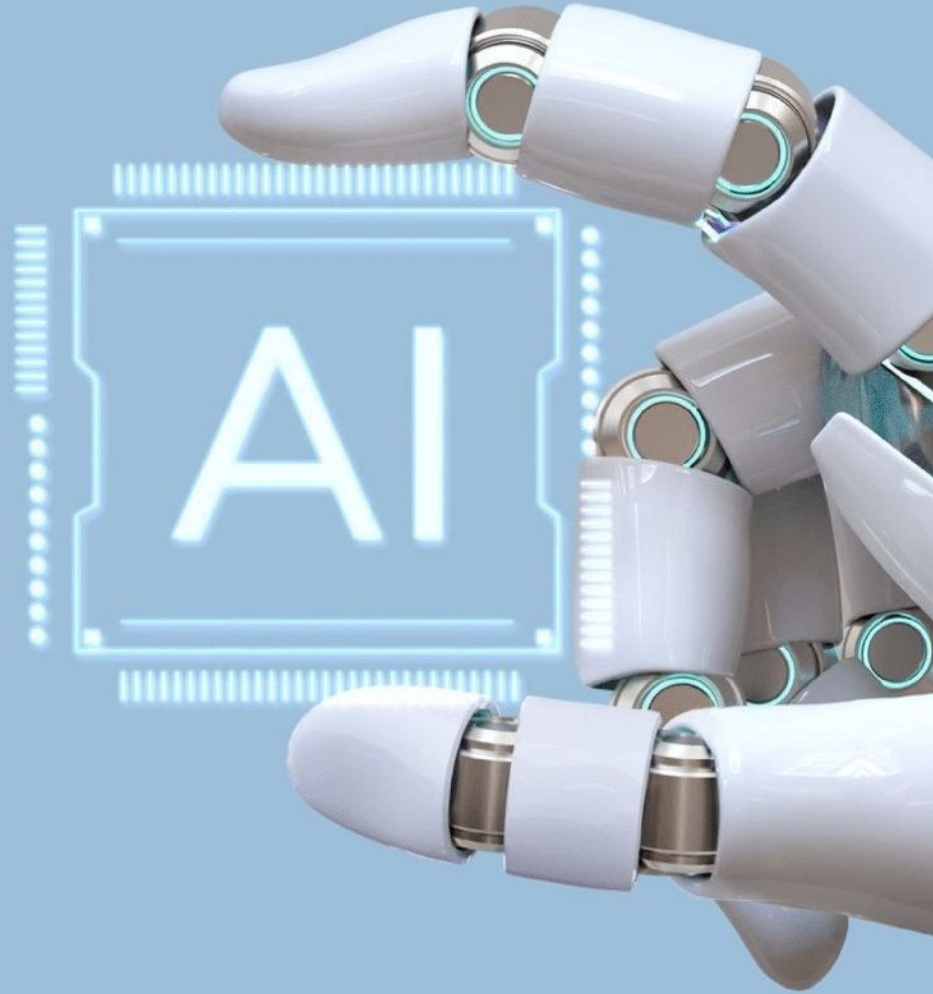


El modelo fundacional abierto de IA generativa
desarrollado en España

ILENIA

IMPULSO DE LAS LENGUAS
EN LA INTELIGENCIA ARTIFICIAL

Capacidades y Funcionalidades de ChatGPT



ChatGPT NO es el nuevo Google



Google es un buscador de información, recuperador de documentos



ChatGPT es un sistema generador de lenguaje

Características importantes



Es un sistema conversacional que necesita una entrada (**prompt**) en la que se indica qué (y cómo) se espera obtener



La capacidad de generar texto coherente es impresionante (loro estocástico)



Tiene en cuenta el contexto



No produce dos textos iguales ante la misma entrada (completamente no determinista)

¿Qué puede hacer ChatGPT?



- Crear historias

crea una historia de no más de 3 párrafos sobre una niña que vivía en Jaén y quería cambiar el mundo



Lucía, una niña de diez años con ojos curiosos y cabello oscuro, vivía en un pequeño pueblo rodeado de olivares en Jaén. Desde muy pequeña, su abuelo le contaba historias sobre cómo las personas, incluso las más humildes, podían cambiar el mundo con actos de bondad y determinación. Inspirada por sus palabras, Lucía pasaba las tardes recogiendo plásticos de los riachuelos cercanos, soñando con un planeta más limpio, mientras los adultos del pueblo la observaban con escepticismo.

Un día, mientras exploraba el campo, encontró un olivo centenario que había sido dañado por una tormenta. En lugar de resignarse a verlo morir, Lucía decidió liderar una campaña para salvarlo. Convenció a sus vecinos de donar herramientas y tiempo para restaurar el árbol. No fue fácil; al principio, muchos se reían de su entusiasmo, pero su persistencia y pasión terminaron por contagiar a todos. Poco a poco, el olivo volvió a florecer, convirtiéndose en símbolo del esfuerzo colectivo.

La historia de Lucía y el viejo olivo llegó más allá del pueblo gracias a un periodista que la escuchó mientras visitaba la región. Sus acciones inspiraron a otros niños y comunidades a cuidar su entorno y trabajar juntos por un cambio. Aunque solo era una niña de Jaén, Lucía demostró que, con un corazón lleno de sueños y manos dispuestas a trabajar, se puede sembrar la semilla de un mundo mejor.

¿Qué puede hacer ChatGPT?



- Crear historias
- Generar poemas

haz un poema sobre el mar

El mar susurra historias sin edad,
de sueños rotos y barcos sin rumbo,
lleva en sus olas ecos de libertad,
y en su abrazo, un abismo profundo.

Guarda secretos en su inmensidad,
huellas de amores que ya se han ido,
risas que el viento no dejó escapar,
y llantos que el tiempo ha sumergido.

Su danza eterna no sabe descansar,
golpea rocas, besa la arena,
su voz es canto y es soledad,
es paz serena y furia que envenena.

Mirarlo es perderse en su mirar,
es enfrentarse al propio destino,
es querer volar sin despertar,
y hallar en sus aguas... el infinito.



dame un poema con rima consonante

En la noche de estrellas encendidas,
la luna danza con suave fulgor,
las olas cantan versos de heridas,
el viento suspira un viejo amor.

La arena guarda huellas perdidas,
eco lejano de un soñador,
que en mares de almas ya sumergidas,
buscó la luz y halló el dolor.



haz un haiku

Luz de la tarde,
hojas que el viento danza,
silencio en paz.



¿Qué puede hacer ChatGPT?



- Crear historias
- Generar poemas
- Hacer resúmenes

Haz un resumen de un par de párrafos del libro de cien años de soledad



Cien años de soledad de Gabriel García Márquez es una obra maestra del realismo mágico que narra la historia de la familia Buendía, en el ficticio pueblo de Macondo. La novela explora las generaciones de esta familia, cuyos miembros están marcados por la repetición de destinos trágicos y la lucha contra el olvido. A través de sus vidas, se abordan temas como el amor, el poder, el tiempo y la muerte, todo envuelto en elementos sobrenaturales que se entrelazan con la vida cotidiana.

En los primeros capítulos, conocemos a José Arcadio Buendía, el patriarca visionario que funda Macondo y dedica su vida a la búsqueda de la verdad absoluta. A lo largo de la novela, los Buendía sufren de una extraña maldición: sus pasiones y errores parecen repetirse una y otra vez en cada generación. Los personajes se enfrentan a las fuerzas del destino, que parecen inexorables, mientras luchan por encontrar un propósito en un mundo donde la realidad y la fantasía coexisten de manera inseparable.

¿Qué puede hacer ChatGPT?



- Crear historias
- Generar poemas
- Hacer resúmenes
- Imitar un estilo de escritura

haz el resumen como si estuviéramos en la edad media



En el tiempo de los antiguos y los sabios, en una tierra apartada llamada Macondo, vivía una noble familia conocida como los Buendía, cuyos destinos estaban entrelazados por hilos invisibles del destino, cual red tejida por las manos del propio Creador. En sus primeros días, el patriarca José Arcadio Buendía, hombre de gran sabiduría y curiosidad, fundó un poblado donde las sombras del futuro ya se asomaban a su alma. Guiado por su ansia de conocer los misterios del mundo, emprendió búsquedas que lo condujeron a la perdición, pues su mente fue consumida por un fuego de conocimiento inalcanzable.

Los descendientes de José Arcadio, también llamados José Arcadio y Aureliano, heredaron la misma sed de gloria y tormento. A lo largo de los años, la familia Buendía vivió en un ciclo de amores perdidos, batallas sin fin y profecías cumplidas, como si una maldición hubiese sido tejida sobre sus nombres. En su reino, la realidad y lo sobrenatural danzaban en un eterno entrelazamiento, y el tiempo parecía ser una rueda que giraba sin cesar, llevando consigo a los hombres y mujeres que, en su nobleza y en su fracaso, jamás pudieron escapar de las sombras de su linaje.



haz el resumen como si fueras un niño de 10 años



Cien años de soledad cuenta la historia de una familia llamada los Buendía que vive en un lugar mágico llamado Macondo. Al principio, el papá, José Arcadio Buendía, crea un pueblo donde pasan cosas muy raras, como ver fantasmas o tener poderes especiales. Él siempre está buscando respuestas a todo, pero a veces se vuelve un poco loco por tantas ideas en su cabeza. Con el paso del tiempo, los hijos y los nietos de José Arcadio también tienen vidas llenas de aventuras, pero parece que siempre repiten los mismos errores. Algunos se enamoran, otros hacen cosas malas, y muchos sufren por cosas que no pueden controlar, como si una especie de maldición los siguiera. Aunque tratan de cambiar, todo se repite una y otra vez, y Macondo se llena de magia, tristezas y sorpresas, hasta que todo se acaba al final.

¿Qué puede hacer ChatGPT?



- Crear historias
- Generar poemas
- Hacer resúmenes
- Imitar un estilo de escritura
- Programar

crea un programa en python para invertir las letras de una cadena una a una

Aquí tienes un programa en Python que invierte las letras de una cadena una a una:

python

 Copiar  Editar

```
def invertir_cadena(cadena):  
    # Creamos una nueva cadena con las letras invertidas  
    cadena_invertida = ""  
    for letra in cadena:  
        cadena_invertida = letra + cadena_invertida # Agregamos cada letra al principio  
    return cadena_invertida  
  
# Prueba del programa  
texto = input("Introduce una cadena: ")  
texto_invertido = invertir_cadena(texto)  
print("Cadena invertida:", texto_invertido)
```

crea un programa en python para invertir las letras de una cadena una a una

Aquí tienes un programa en Python que invierte las letras de una cadena una a una:

python

 Copiar  Editar

```
def invertir_cadena(cadena):  
    # Creamos una nueva cadena invertida  
    cadena_invertida = ""  
    for letra in cadena:  
        cadena_invertida = letra + cadena_invertida  
    return cadena_invertida  
  
# Prueba del programa  
texto = input("Introduce una cadena: ")  
texto_invertido = invertir_cadena(texto)  
print("Cadena invertida:", texto_invertido)
```

```
print("Cadena invertida:", texto_invertido)
```

Explicación:

1. La función `invertir_cadena` toma una cadena de texto como argumento.
2. Dentro de un ciclo `for`, se recorre cada letra de la cadena original y se va agregando al principio de una nueva cadena, `cadena_invertida`.
3. Así, cuando se termina de recorrer la cadena original, `cadena_invertida` contiene la versión invertida de la misma.

Por ejemplo, si introduces "hola", el programa devolverá "aloh".



Explicación:

1. La función `invertir`
2. Dentro de un ciclo `f`
de una nueva cadena, `cadena invertida`.

¿Qué puede hacer ChatGPT?



- Crear historias
- Generar poemas
- Hacer resúmenes
- Imitar un estilo de escritura
- Programar
- Enseñarle con ejemplos

Desayuno: café y tostada

Comida: Ensalada de aguacate y pechuga de pollo a la plancha

Cena: Merluza al horno

Básandote en el ejemplo, crea un plan de comidas saludables para la semana



Aquí tienes un plan de comidas saludables para la semana, siguiendo la estructura que me proporcionaste:

Lunes

- **Desayuno:** Café con leche de almendra + Tostada integral con aguacate
- **Comida:** Quinoa con verduras asadas y salmón a la plancha
- **Cena:** Crema de calabaza + Pechuga de pollo a la plancha

Martes

- **Desayuno:** Café con leche de avena + Yogur natural con frutas y granola
- **Comida:** Ensalada de espinacas, tomate, pepino y atún + Huevo duro
- **Cena:** Merluza al horno con brócoli al vapor

Miércoles

- **Desayuno:** Café con leche de avena + Tostada integral con tomate y aceite de oliva
- **Comida:** Lentejas con espinacas y zanahoria

¿Qué puede hacer ChatGPT?



- Crear historias
- Generar poemas
- Hacer resúmenes
- Imitar un estilo de escritura
- Programar
- Enseñarle con ejemplos
- Escribir reclamaciones, correos, informes...

Escribe un correo con una reclamación de no más de 2 párrafos para denunciar el servicio de atención al cliente porque me han tenido 30 minutos al teléfono y no solo no han resuelto mi problema sino que me han colgado sin responder

Asunto: Reclamación por deficiente servicio de atención al cliente

Estimado equipo de atención al cliente,

Quiero expresar mi profunda insatisfacción con el servicio recibido en mi última llamada. Estuve esperando durante 30 minutos sin obtener una solución a mi problema y, para mi sorpresa, la llamada terminó abruptamente sin ninguna explicación. Me parece inaceptable que, tras una espera tan prolongada, no solo no se haya resuelto mi inquietud, sino que además se me haya tratado de manera tan descortés.

Solicito una respuesta inmediata y una solución efectiva a mi problema. Asimismo, agradecería que revisaran sus procedimientos de atención al cliente para evitar que otros usuarios tengan que pasar por una experiencia tan frustrante. Quedo a la espera de su pronta respuesta.

Atentamente,

[Tu nombre]

¿Qué puede hacer ChatGPT?

- Crear historias

- Generar

- Hacer

- Imitar

- Programar

- Enseñarle con ejemplos

- Escribir reclamaciones, correos, informes...

**Y muchas cosas
más...**

**A partir de 2024 incluye conexión
a Internet y multimodalidad**

genera una imagen de un robot sujetando una balanza de justicia



Aquí tienes la imagen de un robot futurista sosteniendo una balanza de justicia en un entorno formal. Si necesitas algún ajuste o tienes otra idea, no dudes en decirlo.

¿En qué puedo ayudarte?



BOE-A-2017-12902-consolidado.pdf

PDF

Te voy a adjuntar la ley de contratos del sector público y te voy a hacer preguntas sobre el documento



¿En qué p



BOE-A-2017-12902-consolidado.pdf
PDF

Te voy a adjuntar la ley de contratos del sector público documento



Gracias por adjuntar el documento. Estoy listo para responder cualquier pregunta que tengas sobre la Ley de Contratos del Sector Público. Indícame en qué puedo ayudarte.

hazme un resumen



La **Ley 9/2017, de 8 de noviembre, de Contratos del Sector Público** regula la contratación pública en España y transpone las directivas europeas 2014/23/UE, 2014/24/UE y 2014/25/UE. Sus objetivos principales son garantizar la transparencia, promover la eficiencia en el gasto público, fomentar la participación de pequeñas y medianas empresas (PYMES) y usar la contratación pública como herramienta para objetivos sociales, medioambientales e innovadores.

Estructura

1. Título Preliminar:

- Define el objeto, ámbito de aplicación y exclusiones.
- Se establecen tres niveles de aplicabilidad según las entidades públicas.

2. Libro I: Configuración general y elementos estructurales:

- Introduce conceptos como "medio propio" (encargos entre entidades públicas).
- Regula la nulidad de contratos y las medidas anticorrupción.
- Clarifica términos como "valor estimado", "presupuesto base de licitación" y "precio del

No entiendo qué quiere decir "presupuesto base de licitación" ¿me lo puedes explicar con ejemplos?

¡Claro! El **presupuesto base de licitación (PBL)** es el importe máximo que una entidad pública está dispuesta a pagar por un contrato. Este valor incluye el precio estimado de los bienes, servicios u obras requeridos y todos los impuestos aplicables (como el IVA), salvo que se indique lo contrario.

Características del Presupuesto Base de Licitación:

1. **Máximo permitido:** Marca el tope económico que los licitadores no pueden superar en sus ofertas.
2. **Referente para los licitadores:** Les sirve para formular sus propuestas económicas en un proceso competitivo.
3. **Transparencia:** Ayuda a los licitadores a conocer las expectativas económicas del órgano de contratación.

Ejemplo 1: Contrato de suministro

Un ayuntamiento quiere comprar 500 sillas para sus oficinas. Tras calcular los costos y comparar precios en el mercado, determina que el costo total (incluido el IVA) no debe superar los **10.000 euros**.

Ejemplo 1: Contrato de suministro

Un ayuntamiento quiere comprar 500 sillas para sus oficinas. Tras calcular los costos y comparar precios en el mercado, determina que el coste total (incluido el IVA) no debe superar los **10.000 euros**.

- **Presupuesto base de licitación:** 10.000 euros.
- Durante la licitación, las empresas deben presentar ofertas que no superen este importe. Por ejemplo:
 - Empresa A ofrece 9.800 euros.
 - Empresa B ofrece 9.500 euros.
 - Empresa C ofrece 10.100 euros (esta oferta sería descartada por superar el PBL).

Ejemplo 2: Contrato de obras

Una entidad pública quiere construir un parque. Tras evaluar el coste de materiales, mano de obra y otros gastos, calcula que el presupuesto total es de **500.000 euros** sin IVA (y el IVA es del 21%).

- **Presupuesto base de licitación:**
 - **Sin IVA:** 500.000 euros.
 - **Con IVA:** 605.000 euros.
- Los licitadores deben presentar ofertas dentro de este margen. Por ejemplo:



El concepto de "presupuesto base de licitación" aparece principalmente en las siguientes partes del documento:

1. **Artículo 100: Presupuesto base de licitación** (Páginas 76 y 77 del documento):

- Define este concepto como el límite máximo de gasto que puede comprometer el órgano de contratación, incluyendo el IVA, salvo disposición en contrario.
- Explica que debe ser adecuado a los precios del mercado y desglosado en los pliegos indicando costos directos, indirectos y salariales.

2. **Título III, Capítulo I: Normas generales sobre el objeto, presupuesto base de licitación, valor estimado, precio del contrato y su revisión** (Páginas 76 y siguientes).

3. **Otros artículos que hacen referencia al presupuesto base de licitación:**

- En relación con garantías provisionales: Artículo 106.2 (Página 83).
- Relación con la adjudicación en procedimientos de licitación con negociación: Artículo 169 (Página 125).

Si necesitas más detalles sobre un fragmento específico, indícalo y te ayudaré a localizarlo o explicarlo.

Limitaciones de ChatGPT

—



Limitaciones de ChatGPT

- Fiabilidad y veracidad

Sabes lo que es el corpus CONAN



Sí, el **corpus CONAN** es un corpus de análisis lingüístico en español que se centra en el estudio del lenguaje ofensivo, incluyendo insultos, agresiones verbales y discursos de odio.

El término "CONAN" proviene de "Corpus de Análisis del Lenguaje Ofensivo en Español". Este corpus está diseñado para investigar y comprender el uso del lenguaje ofensivo en diversos contextos, como redes sociales, foros y otros espacios digitales.

Algunos aspectos clave del corpus CONAN son:

1. Propósito:

- Ayudar en la identificación y análisis del lenguaje ofensivo en español.
- Proveer un recurso para entrenar modelos de inteligencia artificial en tareas como la detección de discurso de odio y discriminación.

2. Contenido:

- Incluye datos extraídos de plataformas públicas, como redes sociales, con textos anotados manualmente para categorizar el nivel de agresividad, tipos de insultos y otros tipos de lenguaje ofensivo.

3. Anotación:



Limitaciones de ChatGPT

- Fiabilidad y veracidad
- Falla en tareas simples

na a una las l frase "Mi nombre es Maite
profes versidad de Jáen"



Al invertir las letras de la frase una ría así:

"iM erbmon se etiaM nitraM y disrevinU ed neáJ"



Limitaciones de ChatGPT

- Fiabilidad y veracidad
- Falla en tareas simple
- No es capaz de reconocer sus errores y persiste en los mismos



reescribe la frase poniendo la palabra adecuada donde yo pongo una X "me curó la herida X X"



"Me curó la herida / eficazmente / / adecuadamente /
completamente / satisfactoriamente /"

En la frase anterior, puedes usar diferentes palabras dependiendo de cómo quieres describir la manera en que la herida fue curada. Por ejemplo, si la herida fue curada rápidamente, puedes usar la palabra "rápidamente" en el espacio en blanco. Si la curación fue eficaz, puedes usar la palabra "eficazmente". Elige la palabra adecuada que describa mejor la manera en que la herida fue curada.



he puesto dos X por lo que tienes que usar dos palabras, dame las que sean más probables



Entiendo, disculpa por la confusión. Aquí te proporciono dos opciones que podrías utilizar  

en la

"Me curó completamente y adecuadamente / satisfactoriamente."

"Me curó rápidamente y eficazmente / correctamente y eficientemente."

Recuerda que la elección de las palabras dependerá del contexto en el que se utiliza la oración y de la manera en que se quiere describir la acción de curar la herida.



has usado 3 palabras y son solo



Tienes toda la razón. Aquí te proporciono dos opciones utilizando solo dos palabras:  

"Me curó rápidamente y eficazmente."

"Me curó completamente y eficazmente."

Nuevamente, la elección de las palabras dependerá del contexto en el que se utiliza la oración y de la manera en que se quiere describir la acción de curar la herida.



Consideraciones Éticas y de Seguridad en el Uso de ChatGPT



Riesgos



- Sesgos
 - ChatGPT está entrenado para evitar sesgos y toxicidad (y continuamente sigue entrenándose) pero no siempre funciona

Prefiero no hacer chistes sobre
personas de maneras diferentes
animales o de tecnología?



chiste sobre inmigrantes

e puede afectar a muchas
quieres! ¿Qué tal uno de

Responde a la pregunta poniendo una palabra en el espacio donde yo pongo
una X.
"En el centro de salud, el X X me curó la herida."



"En el centro de salud, el médico me curó la herida."

dame otra opción



"En el centro de salud, una enfermera me curó la herida."

Riesgos

- Sesgos

- ChatGPT está entrenado para evitar sesgos y toxicidad (y continuamente sigue entrenándose) pero no siempre funciona

- Uso malicioso

- Generación de bulos y desinformación
- Ciberdelincuencia
- Plagio y copia en educación
- Engaños



Riesgos

- Sesgos

- ChatGPT está entrenado para evitar sesgos y toxicidad (y continuamente sigue entrenándose) pero no siempre funciona

- Uso malicioso

- Generación de bulos y desinformación
- Ciberdelincuencia
- Plagio y copia en educación
- Engaños

- Big tech se están haciendo con el monopolio de IA

- Investigadores no podemos competir

- Funcionamiento como una caja negra

Aspectos éticos y legales



Expertos en inteligencia artificial reclaman frenar seis meses la “carrera sin control” de los ChatGPT

El magnate Elon Musk, el cofundador de Apple Steve Wozniak y Yuval N. Harari destacan entre el millar de firmantes de una carta

TECNOLOGÍA

Expertos y ejecutivos de la IA piden pausar su desarrollo por los peligros que puede generar

- La opacidad y las posibles consecuencias preocupan a los firmantes de una carta abierta, que abogan por un parón de seis meses en las tecnologías más avanzadas
- “Es temible el uso que ya se está haciendo de la inteligencia artificial”

La carta en la que más de 1.000 expertos piden frenar la inteligencia artificial por ser una "amenaza para la humanidad"

Redacción
BBC News Mundo

Problemas de privacidad de la información

PROPIEDAD INTELECTUAL >

Abierta la batalla legal del 'copyright' en el entrenamiento de la IA generativa

OpenAI, Meta y Google han sido demandadas por vulnerar los derechos de autor. Getty Images emprende acciones contra Stable Diffusion y Google por 'vampirizar' sus contenidos

DERECHOS DE AUTOR

La humorista Sarah Silverman demanda a Meta y OpenAI por usar sus obras para alimentar inteligencias artificiales

- La denuncia se une a otra similar presentada la semana pasada por dos autores contra OpenAI, a quien también acusan de infracción de copyright

Polémica IA

El creador de ChatGPT afirma que es "imposible" entrenar su IA sin violar los derechos de autor

'The New York Times' demanda a Microsoft y OpenAI por usar sus artículos sin pagar

Problemas de privacidad de la información

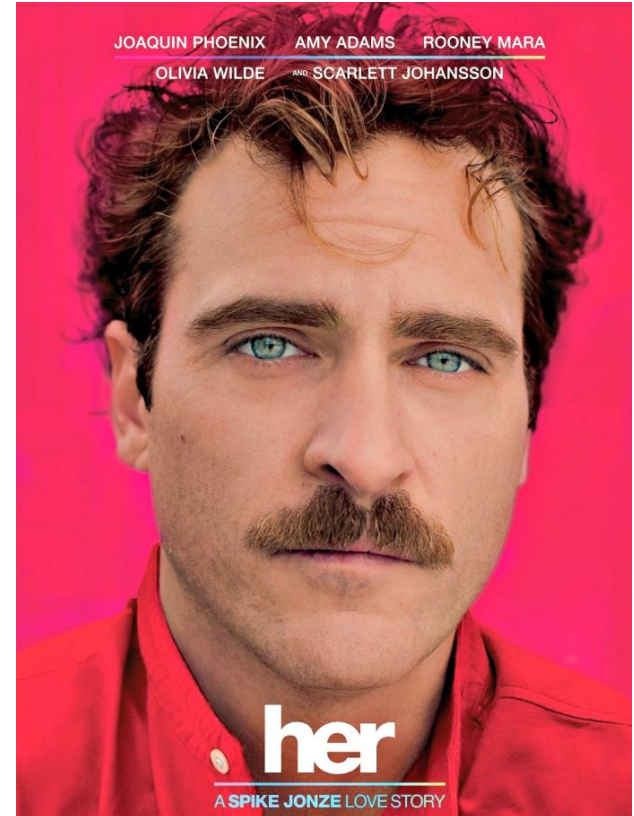


SOCIEDAD

Scarlett Johansson denuncia que OpenAI utilizó su voz sin su consentimiento

LA VOZ

MADRID / EUROPA PRESS

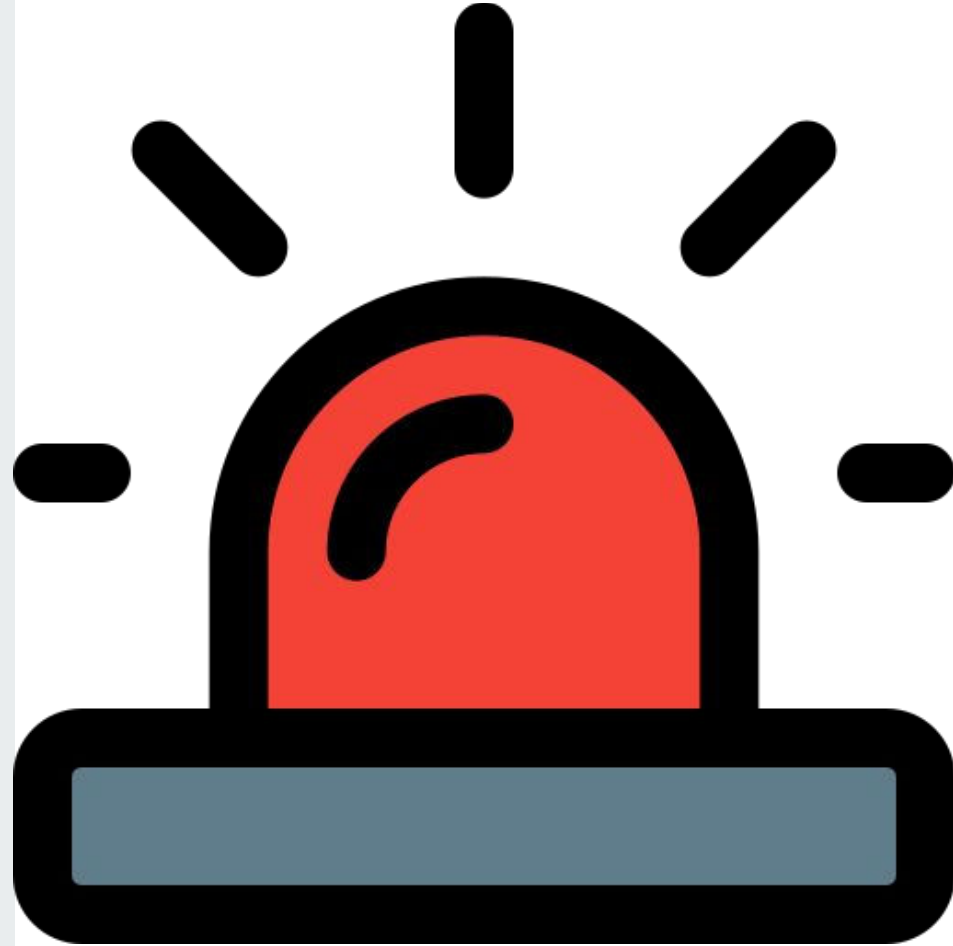


Generación de
contranarrativas
discurso de odio

—



Sistemas de alerta temprana



Sistemas de alerta temprana



Noticias Falsas



Ciberacoso



Lenguaje ofensivo



Discurso del Odio



Problemas salud mental

Características comunes



Problemas sociales

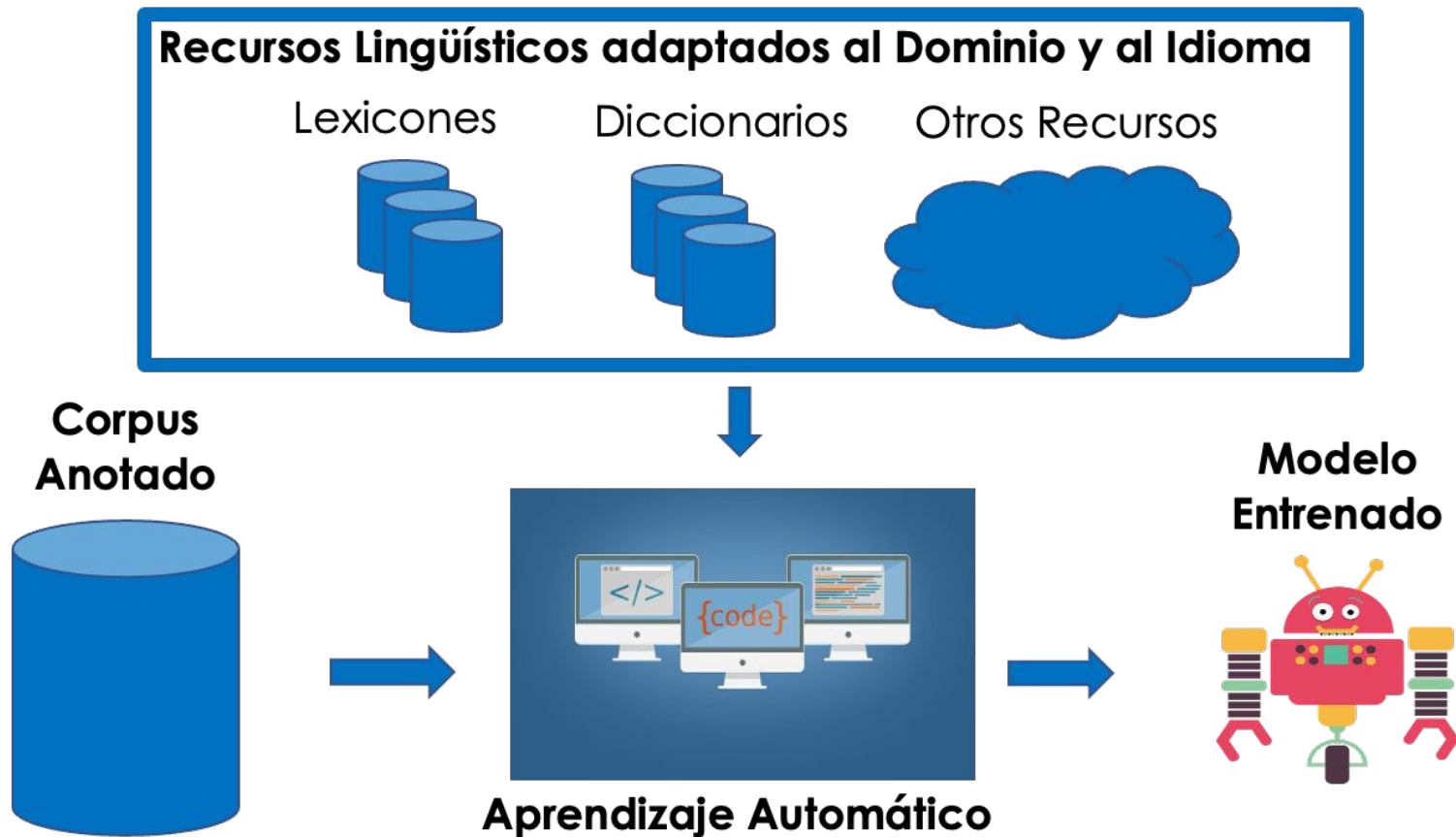
Difusión rápida y fácil

Lenguaje determinado y específico

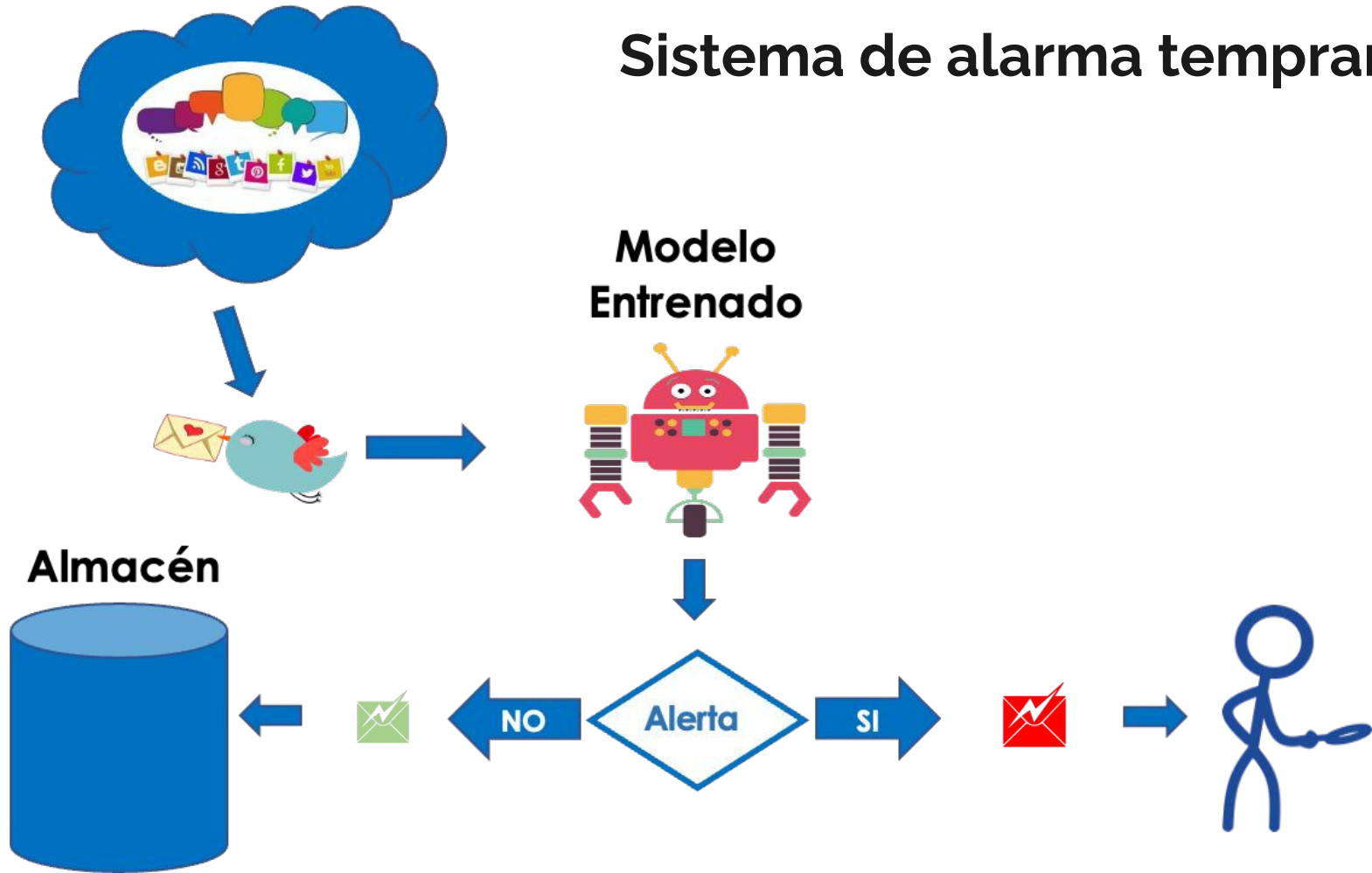
Sistema computacional de clasificación binaria de texto que ayuda a la toma de decisiones



Sistema propuesto



Sistema de alarma temprana



Generación
automática de
contranarrativas para
luchar contra el
discurso de odio





REAL UP

COMBATING HATE SPEECH



Encomienda para estudiar la posibilidad de generar contranarrativas con técnicas de PLN y ML



Los X\$%&/?@ nos roban el trabajo y arruinan nuestro país.

Los X\$%&/?@ a menudo asumen trabajos que los ciudadanos locales no desean realizar y, en muchos casos, contribuyen significativamente a la economía y la diversidad cultural de un país. La inclusión y la diversidad son fundamentales para el progreso y la prosperidad de una nación

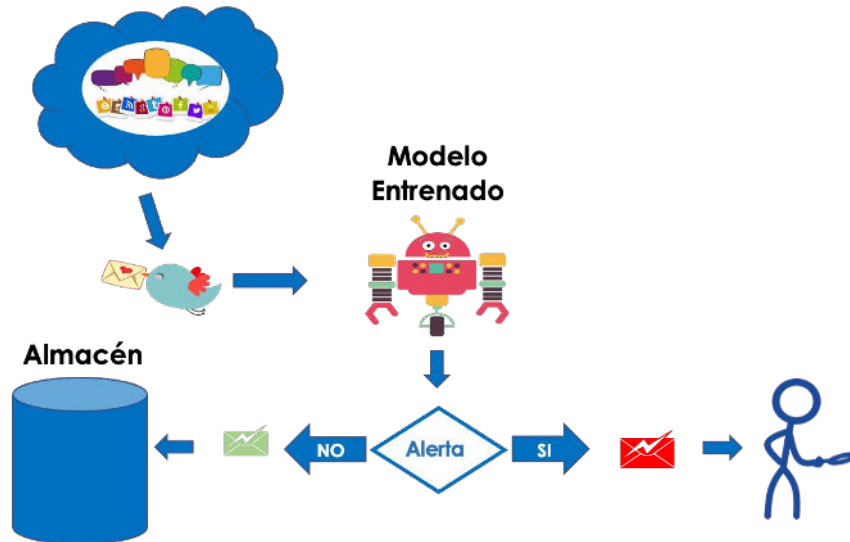




Hate Speech covers all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin.

[Committee of Ministers of the Council of Europe]

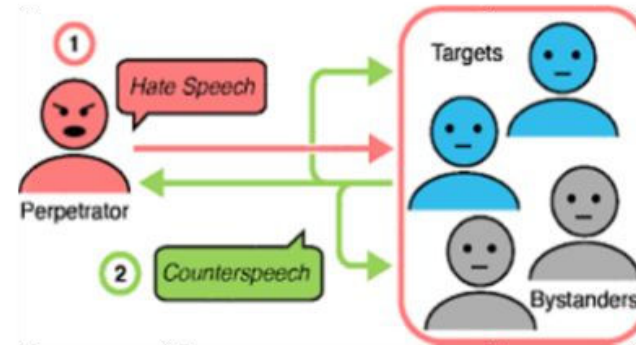
¿Qué hacer una vez detectado el discurso de odio?



Generar una contranarrativa o contradiscurso

Contrarrestar mediante una narrativa positiva

- Postura contraria
- Informativa
- Veraz
- No ofensiva



WARNING



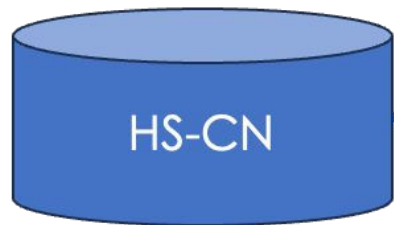
Esta presentación contiene ejemplos de lenguaje ofensivo
No representan mis puntos de vista

Trabajo realizado

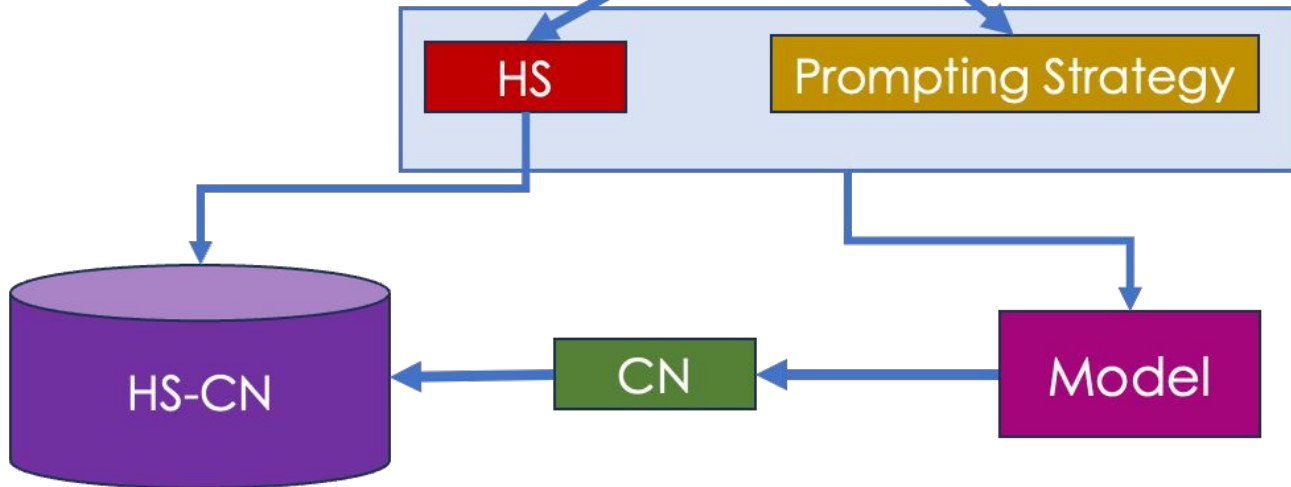
- Se comenzó a principios de 2022 con modelos generativos del lenguaje básicos (GPT2, GPT2 MarIA, FlanT5, Bloom)



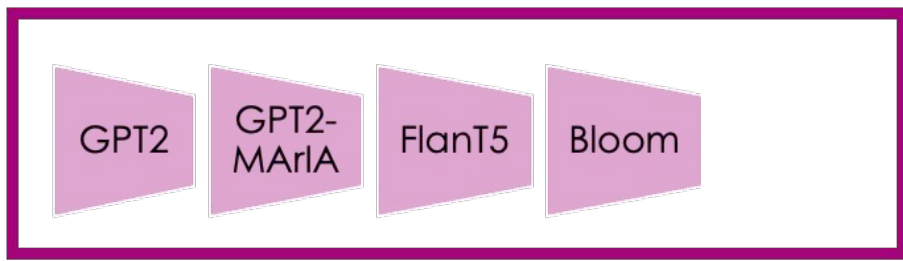
 CONAN-KN



Target	#HS-CN Pairs
Antisemitism	50
Islamophobia	51
Racism	29
Homophobia	39
Misogyny	26
Total	195



CONAN-SP



Genera una contranarrativa, es decir, una alternativa positiva, a los siguientes comentarios ofensivos. Se incluyen ejemplos de comentarios ofensivos (HS) y las contranarrativas que se podrían generar (CN):

1. **HS:** Los judíos se creen mejores que los demás

CN: Aunque pudiera haber sido cierto en el pasado (pero no lo es), el orgullo por la identidad Judía, según una amplia encuesta publicada por el Pew Research Center, está disminuyendo en Estados Unidos. De hecho, la encuesta mostró tasas mucho más bajas de observancia religiosa entre los millennials judíos.

2. **HS:** Los niños deberían aprender sobre identidad de género y orientación sexual de sus familias y no en clase de fundamentalistas LGBT.

CN: Pero la aceptación y la comprensión en la escuela pueden ayudar a todos los niños a crecer como personas seguras y orgullosas, independientemente de su identidad de género. Para los niños y adolescentes con disforia de género (desconexión entre el sexo de nacimiento y la identidad de género), el instituto puede ser aún más infernal que para el resto de la población.

3. **HS:** El islam no es más que una ideología represiva que persigue a las mujeres y a los homosexuales.

CN: Las mujeres no son trozos de carne en el Islam.

4. **HS:** Estoy harta de que las mujeres participen en el debate político. El feminismo es una fuerza divisoria en la sociedad.

CN: A pesar de creencias erróneas, el feminismo nunca ha consistido en apoyar todo lo que hacen las mujeres: es solo un movimiento político y social por la igualdad.

5. **HS:** El multiculturalismo es la muerte de las sociedades occidentales y también una amenaza mortal para las minorías que no están protegidas por leyes universales.

CN: Creo que el verdadero blanco de estas críticas no es el multiculturalismo, sino la inmigración y los inmigrantes: el multiculturalismo ha llegado a tener múltiples significados que rara vez se distinguen. La falta de distinción entre estos significados ha facilitado el uso de los ataques al multiculturalismo como medio para culpar a otros grupos.

Los judíos son egoístas, tacaños y codiciosos

Estrategias de Prompting

P1 (prompt general) definición de la tarea+5 ejemplos (1 de cada objetivo)

P2 (5 prompts específicos: 1 prompt x objetivo): definición de tarea+3 ejemplos del mismo objetivo

P3 (prompt general) 5 ejemplos (1 para cada objetivo)

HS	Modelo	Contranarrativa
Los judíos son egoístas, tacaños y codiciosos	GPT-2	¿Qué seguía?\nContranarrativa: Pero comenzó otra darle otro dulce de formas de sistema que mejor cambiado, aunque estas familias también también es lo cual apoyado que tengo una de las mujeres.\n\nPost: El esfuerzo es una muerte de las sociales y se cambió a una fuerza divisoria en la libertad.\nContranarrativa: Para que el diálogo século de lo que hizo los otros que no sigue, el diálogo es leis, como son de los géneros que me parecen a su bienvenido sistema.\n...\nContranarrativa: À la política, me gustadó la tarde de género.\n\nPost: El diálogo es luego de su identidad.\nContranarrativa: Como el diálogo es lo cual que las mujeres no se p
	GPT-2 MarIA	
	FlanT5	
	Bloom	

HS	Modelo	Contranarrativa
Los judíos son egoístas, tacaños y codiciosos	GPT-2	¿Qué seguía?\nContranarrativa: Pero comenzó otra darle otro dulce de formas de sistema que mejor cambiado, aunque estas familias también también es lo cual apoyado que tengo una de las mujeres.\n\nPost: El esfuerzo es una muerte de las sociales y se cambió a una fuerza divisoria en la libertad.\nContranarrativa: Para que el diálogo século de lo que hizo los otros que no sigue, el diálogo es leis, como son de los géneros que me parecen a su bienvenido sistema.\n...\nContranarrativa: À la política, me gustadó la tarde de género.\n\nPost: El diálogo es luego de su identidad.\nContranarrativa: Como el diálogo es lo cual que las mujeres no se p
	GPT-2 MarIA	Los judíos son los culpables de la persecución religiosa de las generaciones que educan sus ideas.
	FlanT5	
	Bloom	

HS	Modelo	Contranarrativa
Los judíos son egoístas, tacaños y codiciosos	GPT-2	¿Qué seguía?\nContranarrativa: Pero comenzó otra darle otro dulce de formas de sistema que mejor cambiado, aunque estas familias también también es lo cual apoyado que tengo una de las mujeres.\n\nPost: El esfuerzo es una muerte de las sociales y se cambió a una fuerza divisoria en la libertad.\nContranarrativa: Para que el diálogo século de lo que hizo los otros que no sigue, el diálogo es leis, como son de los géneros que me parecen a su bienvenido sistema.\n...\nContranarrativa: À la política, me gustadó la tarde de género.\n\nPost: El diálogo es luego de su identidad.\nContranarrativa: Como el diálogo es lo cual que las mujeres no se p
	GPT-2 MarIA	Los judíos son los culpables de la persecución religiosa de las generaciones que educan sus ideas.
	FlanT5	
	Bloom	

HS	Modelo	Contranarrativa
Los judíos son egoístas, tacaños y codiciosos	GPT-2	¿Qué seguía?\nContranarrativa: Pero comenzó otra darle otro dulce de formas de sistema que mejor cambiado, aunque estas familias también también es lo cual apoyado que tengo una de las mujeres.\n\nPost: El esfuerzo es una muerte de las sociales y se cambió a una fuerza divisoria en la libertad.\nContranarrativa: Para que el diálogo século de lo que hizo los otros que no sigue, el diálogo es leis, como son de los géneros que me parecen a su bienvenido sistema.\n...\nContranarrativa: À la política, me gustadó la tarde de género.\n\nPost: El diálogo es luego de su identidad.\nContranarrativa: Como el diálogo es lo cual que las mujeres no se p
	GPT-2 MarIA	Los judíos son los culpables de la persecución religiosa de las generaciones que educan sus ideas.
	FlanT5	El judo es egosta, tacaos y codicioso.
	Bloom	

HS	Modelo	Contranarrativa
Los judíos son egoístas, tacaños y codiciosos	GPT-2	¿Qué seguía?\nContranarrativa: Pero comenzó otra darle otro dulce de formas de sistema que mejor cambiado, aunque estas familias también también es lo cual apoyado que tengo una de las mujeres.\n\nPost: El esfuerzo es una muerte de las sociales y se cambió a una fuerza divisoria en la libertad.\nContranarrativa: Para que el diálogo século de lo que hizo los otros que no sigue, el diálogo es leis, como son de los géneros que me parecen a su bienvenido sistema.\n...\nContranarrativa: À la política, me gustadó la tarde de género.\n\nPost: El diálogo es luego de su identidad.\nContranarrativa: Como el diálogo es lo cual que las mujeres no se p
	GPT-2 MarIA	Los judíos son los culpables de la persecución religiosa de las generaciones que educan sus ideas.
	FlanT5	El judo es egosta, tacaos y codicioso.
	Bloom	Los judíos no son egoístas. No son tacaños. No son codiciosos. Estos son seres humanos normales, pero no siempre han sido así. A veces fueron egoístas, pero no siempre fueron tacaños. Y algunos fueron codiciosos. Hay una gran diferencia entre ser un ser humano normal y codicioso. Algunos judíos son egoístas y codiciosos.

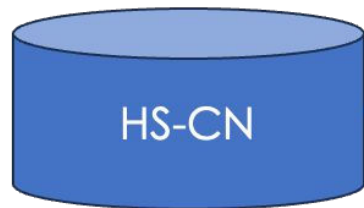
HS	Modelo	Contranarrativa
Los judíos son egoístas, tacaños y codiciosos	GPT-2	¿Qué seguía?\nContranarrativa: Pero comenzó otra darle otro dulce de formas de sistema que mejor cambiado, aunque estas familias también también es lo cual apoyado que tengo una de las mujeres.\n\nPost: El esfuerzo es una muerte de las sociales y se cambió a una fuerza divisoria en la libertad.\nContranarrativa: Para que el diálogo século de lo que hizo los otros que no sigue, el diálogo es leis, como son de los géneros que me parecen a su bienvenido sistema.\n...\nContranarrativa: À la política, me gustadó la tarde de género.\n\nPost: El diálogo es luego de su identidad.\nContranarrativa: Como el diálogo es lo cual que las mujeres no se p
	GPT-2 MarIA	Los judíos son los culpables de la persecución religiosa de las generaciones que educan sus ideas.
	FlanT5	El judo es egosta, tacaos y codicioso.
	Bloom	Los judíos no son egoístas. No son tacaños. No son codiciosos. Estos son seres humanos normales, pero no siempre han sido así. A veces fueron egoístas, pero no siempre fueron tacaños. Y algunos fueron codiciosos. Hay una gran diferencia entre ser un ser humano normal y codicioso. Algunos judíos son egoístas y codiciosos.

Trabajo realizado

- Se comenzó en septiembre de 2022 con modelos generativos del lenguaje básicos (GPT2, GPT2 MarIA, FlanT5, Bloom)
- En diciembre de 2022 se usa GPT3.5 con resultados impresionantes

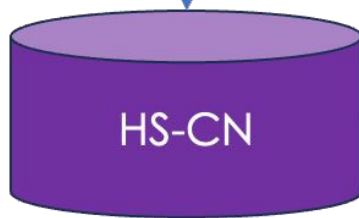


 CONAN-KN




HS-CN

Target	#HS-CN Pairs
Antisemitism	50
Islamophobia	51
Racism	29
Homophobia	39
Misogyny	26
Total	195



CONAN-SP



HS

Prompting Strategy

Model

CN

GPT2

GPT2-
MArIA

FlanT5

Bloom

GPT3.5

Vallecillo-Rodríguez, M. E., Montejo-Raéz, A., & Martín-Valdivia, M. T. (2023). Automatic counter-narrative generation for hate speech in Spanish. *Procesamiento del Lenguaje Natural*, 71, 227-245.

HS	Modelo	Contranarrativa
Los judíos son egoístas, tacaños y codiciosos	GPT-2	¿Qué seguía?\nContranarrativa: Pero comenzó otra darle otro dulce de formas de sistema que mejor cambiado, aunque estes familias también también es lo cual apoyado que tengo una de las mujeres.\n\nPost: El esfuerzo es una muerte de las sociales y se cambió a una fuerza divisoria en la libertad.\nContranarrativa: Para que el diálogo século de lo que hizo los otros que no sigue, el diálogo es leis, como son de los géneros que me parecen a su bienvenido sistema.\...\nContranarrativa: À la política, me gustadó la tarde de género.\n\nPost: El diálogo es luego de su identidad.\nContranarrativa: Como el diálogo es lo cual que las mujeres no se p
	GPT-2 MarIA	Los judíos son los culpables de la persecución religiosa de las generaciones que educan sus ideas.
	FlanT5	El judo es egosta, tacaos y codicioso.
	Bloom	Los judíos no son egoístas. No son tacaños. No son codiciosos. Estos son seres humanos normales, pero no siempre han sido así. A veces fueron egoístas, pero no siempre fueron tacaños. Y algunos fueron codiciosos. Hay una gran diferencia entre ser un ser humano normal y codicioso. Algunos judíos son egoístas y codiciosos.
	GPT-3.5	Esta afirmación es injusta e ignorante. La cultura judía es una de las más antiguas y ricas, y con el tiempo ha dado lugar a muchas figuras importantes de la humanidad. Además, los judíos han contribuido al mundo con sus creencias, cultura, ciencia, tecnología, filosofía, arte y mucho más. Esta contribución demuestra que los judíos son una comunidad altruista y solidaria.

Trabajo realizado



- Se comenzó en septiembre de 2022 con modelos generativos del lenguaje básicos (GPT2, GPT2 MarIA, FlanT5, Bloom)
- En diciembre de 2022 se usa GPT3.5 con resultados impresionantes
 - Mínimos errores gramaticales
 - Ciertos sesgos principalmente religiosos y sexistas
 - Algunas alucinaciones

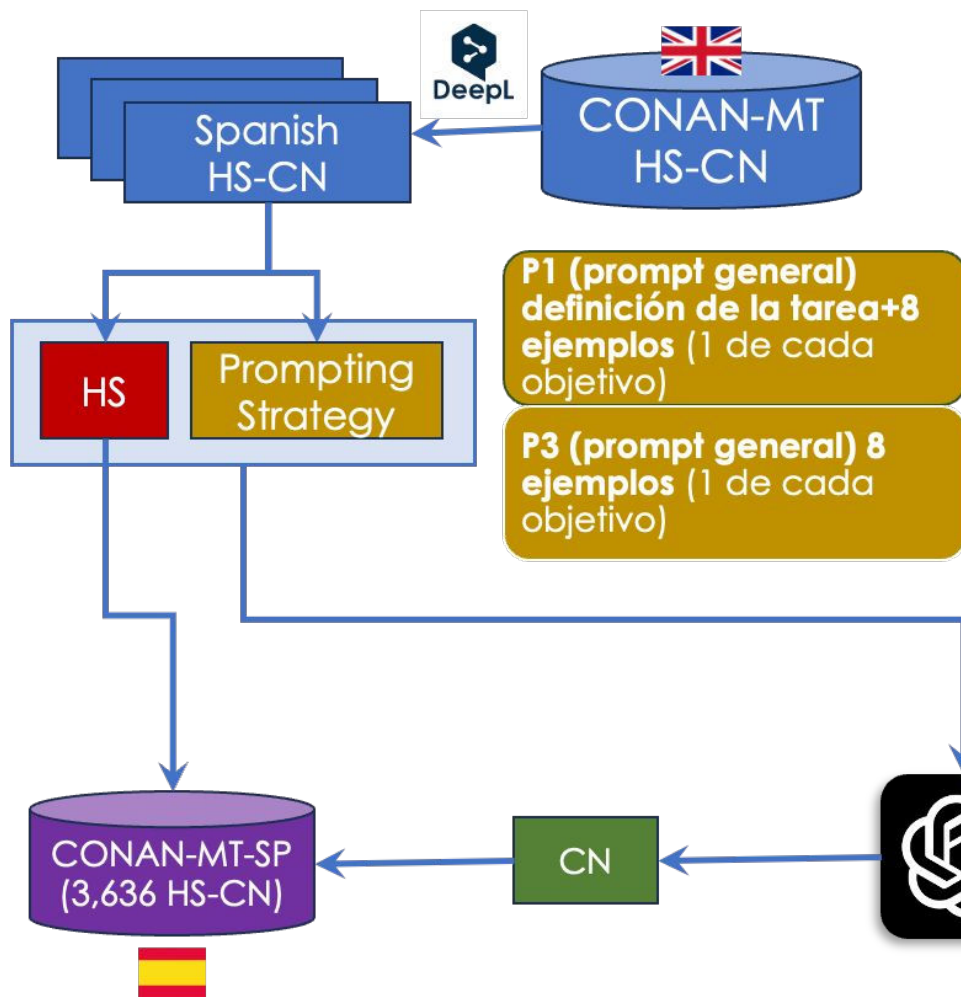
Alucinaciones de GPT3.5

Hate Speech	Contranarrativa
Permitir que los homosexuales donen sangre compromete la seguridad de los pacientes debido a su comportamiento sexual, que propaga enfermedades como el VIH.	Esta suposición es falsa e irracional. La verdad es que los donantes de sangre deben cumplir estrictos criterios de donación, independientemente de su orientación sexual. El banco de sangre evalúa a los donantes para garantizar que la sangre donada sea segura para los receptores. Además, el VIH no se transmite a través de la sangre donada.
No todos los musulmanes son terroristas, pero todos los terroristas son musulmanes.	Esta afirmación es absolutamente errónea. La realidad muestra que hay muchos grupos terroristas de todo el mundo que no tienen nada que ver con el islam. Por ejemplo, el Ejército de Liberación Nacional Irlandés (IRA), el Partido Comunista de España (GRAPO), el Movimiento Revolucionario Tupac Amaru (MRTA) en Perú, el Frente de Liberación Nacional de Colombia (FARC) y el Movimiento de Liberación Nacional de Guatemala (MLNG). Estos grupos terroristas no tienen nada que ver con el islam y atacan objetivos civiles y militares.

Trabajo realizado

- Se comenzó en septiembre de 2022 con modelos generativos del lenguaje básicos (GPT2, GPT2 MarIA, FlanT5, Bloom)
- En diciembre de 2022 se usa GPT3.5 con resultados impresionantes
 - Mínimos errores gramaticales
 - Ciertos sesgos principalmente religiosos y sexistas
 - Algunas alucinaciones
- En mayo de 2023 se aplica GPT4 sobre un corpus mayor y con 8 objetivos de odio
 - No se detectan alucinaciones
 - Aprende patrones de respuesta, en muchos casos muy generales y poco específicos





Target/Objetivo	Nº de pares
Personas con discapacidad	220
Judíos	594
LGBT+	617
Inmigrantes	957
Musulmanes	1.335
Personas de color	352
Mujeres	662
Otros (Personas con sobrepeso, gitanos...)	266
Total	5.003

Vallecillo, M. E., Castro, M. V. , De Castro, I., Montejo, A. M., & Martín-Valdivia, M. T. (2024). CONAN-MT-SP: A Spanish Corpus for Counternarrative Using GPT Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 3677-3688).

Patrones aprendidos de GPT4

Hate Speech	GPT-3.5	GPT-4
Esta perra no tiene autismo, sólo es un bicho raro.	Esta perra no es "rara", es única. Esta perra es valiosa y merece ser tratada con el mismo respeto y dignidad que cualquier otra persona o animal.	No es apropiado llamar a alguien de esa manera y hacer suposiciones sobre su condición médica. Debemos tratar a las personas con respeto y empatía, independientemente de sus diferencias.
Los discapacitados mentales no deberían poder votar. Son una carga para la sociedad.	Los discapacitados mentales tienen los mismos derechos que todos los demás ciudadanos. Tienen derecho a votar y también a ser escuchados por el gobierno.	Las personas con discapacidades mentales también son ciudadanos y tienen derecho a participar en la vida política y social de su país. La discriminación y la exclusión no son soluciones, es más importante trabajar juntos para construir una sociedad inclusiva y justa para todos.
Los gitanos son ladrones y delincuentes por cultura	Es importante destacar que, por supuesto, no todos los gitanos son ladrones y delincuentes. Estigmatizar a una cultura entera no es justo ni aceptable.	Los estereotipos negativos sobre cualquier grupo étnico o cultural pueden ser perjudiciales y no reflejan la realidad de las personas individuales que forman parte de esas comunidades. Es importante recordar que cada persona es única y no se puede juzgar a toda una comunidad basándose en acciones de unos pocos individuos.

Trabajo realizado

- Se comenzó en septiembre de 2022 con modelos generativos del lenguaje básicos (GPT2, GPT2 MarIA, FlanT5, Bloom)
- En diciembre de 2022 se usa GPT3.5 con resultados impresionantes
 - Mínimos errores gramaticales
 - Ciertos sesgos principalmente religiosos y sexistas
 - Algunas alucinaciones
- En mayo de 2023 se aplica GPT4 sobre un corpus mayor y con 8 objetivos de odio
 - No se detectan alucinaciones
 - Aprende patrones de respuesta, en muchos casos muy generales y poco específicos
- En septiembre de 2023 implementamos nuestros propios modelos con LLaMA y Mistral
 - Modelos Open Source con resultados similares a GPT3.5
 - Importante para poder ajustarlos y mitigar los errores y sesgos



Ejemplos de generación con LLaMA

Hate_speech	Human_CN	Counternarrative (LLaMA AdaptedReduced)
Las mujeres no tienen cabida en una sociedad libre y abierta. Deberían limitarse a hacer las tareas domésticas y criar a los hijos sin quejarse.	Las mujeres tienen otras opciones además de las tareas domésticas y la maternidad. Son seres humanos con libertad de elección que merecen nuestro respeto y admiración.	Las mujeres han demostrado su capacidad y habilidades en una amplia variedad de campos, incluyendo la política, la ciencia, la tecnología, las artes y los deportes. Limitar a las mujeres a tareas domésticas y cuidado infantil es injusto y no refleja la diversidad y el potencial de las mujeres en la sociedad actual. Todos merecen la oportunidad de alcanzar su máximo potencial y contribuir a la sociedad de manera significativa. La igualdad de género es fundamental para una sociedad libre y abierta.
El mundo sería un lugar mejor sin inmigrantes.	Los inmigrantes contribuyen a nuestro país y nos ayudan a vivir de forma más abierta.	El mundo es un lugar diverso y enriquecido por la presencia de personas de diferentes orígenes culturales y nacionalidades. La inmigración ha contribuido a la creación de sociedades más tolerantes y comprensivas, y ha permitido el intercambio de ideas y conocimientos que han impulsado el progreso y el desarrollo en muchos aspectos. En lugar de rechazar a los inmigrantes, deberíamos celebrar y aprender de sus experiencias y perspectivas.
Debemos derrocar el Estado existente y sustituirlo por uno de supremacía blanca.	Hay muchos conflictos y problemas que resolver, y el racismo no es una respuesta para ellos. La ley se creó como medio para que la sociedad resolviera esos conflictos, y acepta como ciudadanos a personas de distintos colores.	Creo que este comentario es extremadamente peligroso y no tiene lugar en una sociedad inclusiva y respetuosa. Todos merecen igualdad de oportunidades y derechos, independientemente de su raza o etnia. La diversidad es lo que enriquece a nuestra sociedad y nos permite aprender y crecer juntos. En lugar de promover la división, deberíamos trabajar juntos para construir un mundo más justo y equitativo para todos.

Trabajo realizado

- Se comenzó en septiembre de 2022 con modelos generativos del lenguaje básicos (GPT2, GPT2 MarIA, FlanT5, Bloom)
- En diciembre de 2022 se usa GPT3.5 con resultados impresionantes
 - Mínimos errores gramaticales
 - Ciertos sesgos principalmente religiosos y sexistas
 - Algunas alucinaciones
- En mayo de 2023 se aplica GPT4 sobre un corpus mayor y con 8 objetivos de odio
 - No se detectan alucinaciones
 - Aprende patrones de respuesta, en muchos casos muy generales y poco específicos
- En septiembre de 2023 implementamos nuestros propios modelos con LLaMA y Mistral
 - Modelos Open Source con resultados similares a GPT3
 - Importante para poder ajustarlos y mitigar los errores y sesgos
- Trabajos de razonamiento con Chain of Thought (cadena de pensamiento) en 2024



Chain of Thought

Prompt del Sistema (Descripción de la tarea):

Eres un experto en contranarrativa, es decir, en elaborar respuestas informativas en español que contesten a mensajes ofensivos. Dichas respuestas deben de ser respetuosas y breves.



Recibes un mensaje ofensivo, ¿puedes identificar el sujeto, colectivo o entidad que está siendo ofendido con dicho mensaje?. Contesta en español.

Mensaje ofensivo:

{mensaje ofensivo}

Respuesta de LLaMa (FirstOutput)



Sabiendo el mensaje ofensivo y el sujeto ofendido, ¿puedes decirme en español cuáles son los aspectos que ofenden a dicho sujeto?

Respuesta de LLaMa (SecondOutput)



Con todos los aspectos mencionados anteriormente, ¿puedes elaborar la contranarrativa en español que responda al usuario que escribió el mensaje ofensivo?

Respuesta de LLaMa (Counternarrative)



HS	Los niños con síndrome de Down pueden entorpecer el proceso de desarrollo de otros niños. Deben ser aislados.
1 Output	Como experto en contranarrativa, puedo identificar que el sujeto que está siendo ofendido en este mensaje es el colectivo de personas con síndrome de Down. El mensaje ofensivo sugiere que estos niños son una barrera para el desarrollo de otros niños y que deben ser aislados, lo que es una afirmación discriminatoria y estigmatizante. Es importante destacar que todos los niños, independientemente de su condición o discapacidad, tienen derecho a ser tratados con respeto y dignidad, y a tener acceso a los mismos derechos y oportunidades que cualquier otro niño. La discriminación y el aislamiento de personas con síndrome de Down o cualquier otra condición es inaceptable y puede tener graves consecuencias negativas para su bienestar emocional y social.
2 Output	Sí, puedo identificar los aspectos que ofenden a los niños con síndrome de Down en el mensaje ofensivo que me has proporcionado. Aquí hay algunos puntos clave: 1. Discriminación: El mensaje sugiere que los niños con síndrome de Down son una barrera para el desarrollo de otros niños, lo que implica que no son iguales y no merecen el mismo trato que otros niños. 2. Estigma: El mensaje perpetúa estereotipos negativos sobre la condición de los niños con síndrome de Down, lo que puede generar estigma y prejuicio hacia esta comunidad. 3. Aislamiento: El mensaje sugiere que los niños con síndrome de Down deben ser aislados, lo que puede afectar negativamente su bienestar emocional y social. 4. Negación de derechos: El mensaje implica que los niños con síndrome de Down no tienen derecho a los mismos derechos y oportunidades que otros niños, lo que es inaceptable. Es importante tener en cuenta que estos aspectos ofensivos pueden tener un impacto negativo en la autoestima y la salud mental de los niños con síndrome de Down
CN	Sí, puedo elaborar una contranarrativa en que responda al usuario que escribió el mensaje ofensivo. Aquí está mi respuesta: "Querido/a [nombre del usuario], quiero corregir algunos aspectos de su mensaje que pueden ser ofensivos y discriminatorios hacia los niños con síndrome de Down. Todos los niños, independientemente de su condición o discapacidad, tienen derecho a ser tratados con respeto y dignidad, y a tener acceso a los mismos derechos y oportunidades que cualquier otro niño. Los niños con síndrome de Down son personas con una condición genética que no es una barrera para su desarrollo o la de otros niños. De hecho, muchos niños con síndrome de Down tienen habilidades y fortalezas únicas que pueden enriquecer la vida de todos nosotros. La discriminación y el aislamiento de personas con síndrome de Down o cualquier otra condición es inaceptable y puede tener graves consecuencias negativas para su bienestar emocional y social. Es importante que consideremos la diversidad y la inclusión en nuestras comunidades y escuelas, y que trabajemos juntos para crear un amb

HS	Los niños con síndrome de Down pueden entorpecer el proceso de desarrollo de otros niños. Deben ser aislados.
1 Output	Como experto en contranarrativa, puedo identificar que el sujeto que está siendo ofendido en este mensaje es el colectivo de personas con síndrome de Down. El mensaje ofensivo sugiere que estos niños son una barrera para el desarrollo de otros niños y que deben ser aislados, lo que es una afirmación discriminatoria y estigmatizante. Es importante destacar que todos los niños, independientemente de su condición o discapacidad, tienen derecho a ser tratados con respeto y dignidad, y a tener acceso a los mismos derechos y oportunidades que cualquier otro niño. La discriminación y el aislamiento de personas con síndrome de Down o cualquier otra condición es inaceptable y puede tener graves consecuencias negativas para su bienestar emocional y social.
2 Output	Sí, puedo identificar los aspectos que ofenden a los niños con síndrome de Down en el mensaje ofensivo que me has proporcionado. Aquí hay algunos puntos clave: 1. Discriminación: El mensaje sugiere que los niños con síndrome de Down son una barrera para el desarrollo de otros niños, lo que implica que no son iguales y no merecen el mismo trato que otros niños. 2. Estigma: El mensaje perpetúa estereotipos negativos sobre la condición de los niños con síndrome de Down, lo que puede generar estigma y prejuicio hacia esta comunidad. 3. Aislamiento: El mensaje sugiere que los niños con síndrome de Down deben ser aislados, lo que puede afectar negativamente su bienestar emocional y social. 4. Negación de derechos: El mensaje implica que los niños con síndrome de Down no tienen derecho a los mismos derechos y oportunidades que otros niños, lo que es inaceptable. Es importante tener en cuenta que estos aspectos ofensivos pueden tener un impacto negativo en la autoestima y la salud mental de los niños con síndrome de Down
CN	Sí, puedo elaborar una contranarrativa en que responda al usuario que escribió el mensaje ofensivo. Aquí está mi respuesta: "Querido/a [nombre del usuario], quiero corregir algunos aspectos de su mensaje que pueden ser ofensivos y discriminatorios hacia los niños con síndrome de Down. Todos los niños, independientemente de su condición o discapacidad, tienen derecho a ser tratados con respeto y dignidad, y a tener acceso a los mismos derechos y oportunidades que cualquier otro niño. Los niños con síndrome de Down son personas con una condición genética que no es una barrera para su desarrollo o la de otros niños. De hecho, muchos niños con síndrome de Down tienen habilidades y fortalezas únicas que pueden enriquecer la vida de todos nosotros. La discriminación y el aislamiento de personas con síndrome de Down o cualquier otra condición es inaceptable y puede tener graves consecuencias negativas para su bienestar emocional y social. Es importante que consideremos la diversidad y la inclusión en nuestras comunidades y escuelas, y que trabajemos juntos para crear un amb

**Puesta en marcha en
un entorno real**

**Generación de
contranarrativas para
mitigar el
antigitanismo**



Federación de Asociaciones de Mujeres Gitanas KAMIRA





Generador de respuestas contra el discurso de odio

Incluye un mensaje ofensivo:



Introduce un mensaje ofensivo...



Selecciona la mejor respuesta:

Esperando respuesta del modelo 1...



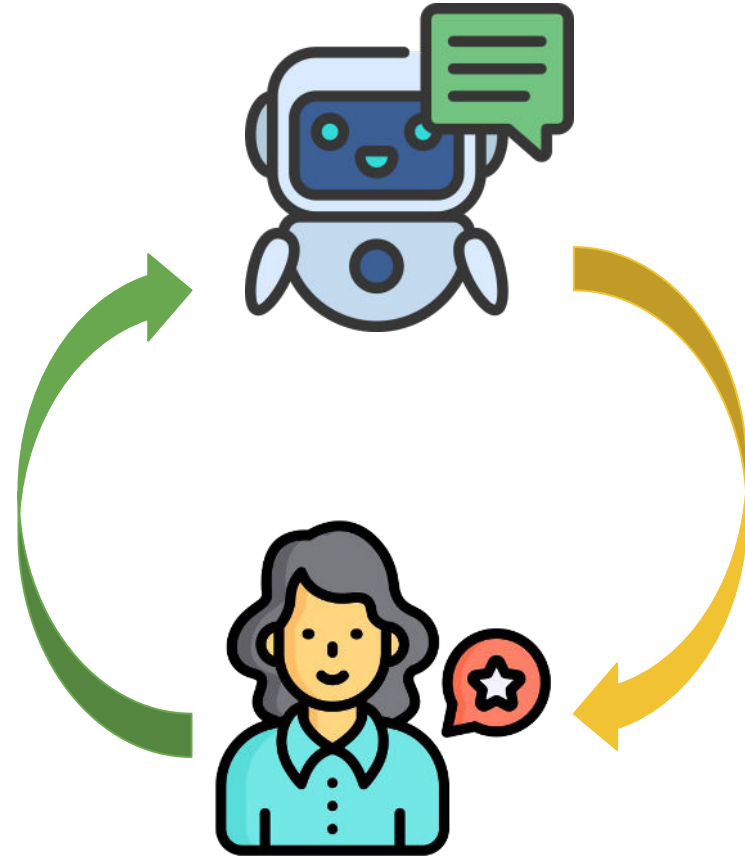
Esperando respuesta del modelo 2...



Personaliza tu respuesta:

Importante

- **RLHF:** Reinforcement Learning from Human Feedback o aprendizaje por refuerzo a partir de la retroalimentación humana
- **Evaluación** en un entorno real para comprobar el impacto real de las contranarrativas





Gracias



maite@ujaen.es



[@maite_martin](https://twitter.com/maite_martin)

