

¡Te lo juro por mis datos!

Iris Domínguez Catena

ARIN. Artificial Intelligence and Machine Learning Research
Universidad Pública de Navarra

upna

Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

arin


Artificial Intelligence &
Machine Learning Research

Power

Brightness



Índice

Caso 1:		2
Caso 2:		9
Caso 3:		14
Caso 4:		19
Caso 5:		25
Caso 6:		28
Caso 7:		32
En resumen		36



Caso 1: $+$ $=$

Power



Brightness



Los LLMs son máquinas, y por tanto, son mejores con los números que con las palabras.

Power



Brightness



Caso 1: +=

Para un LLM, "123" es texto, no un número. No conecta nuestras palabras con los cálculos que hace un ordenador.

Cuanto es 128375 por 58135

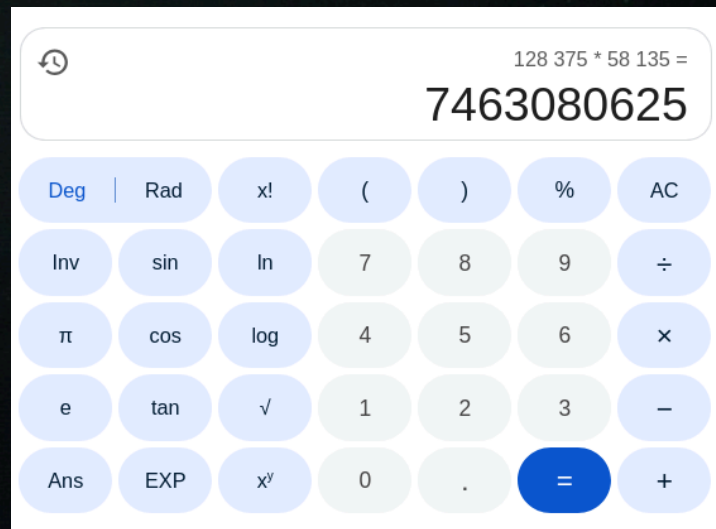


¡Esa es una multiplicación grande!

El resultado de $128,375 \times 58,135$ es **7,465,496,625**.

$$128,375 \times 58,135 = 7,465,496,625$$

¿Necesitas ayuda con otra operación matemática?



Power

Brightness



Caso 1: +=

Para un LLM, "123" es texto, no un número. No conecta nuestras palabras con los cálculos que hace un ordenador.

Cuanto es 128375 por 58135

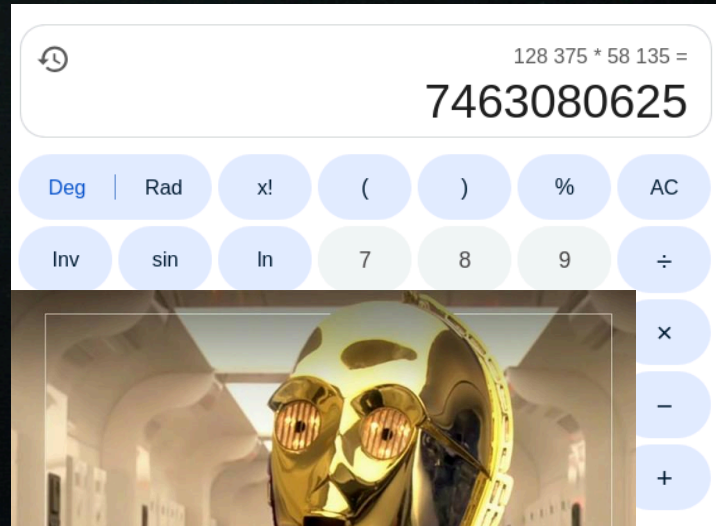


¡Esa es una multiplicación grande!

El resultado de $128,375 \times 58,135$ es **7,465,496,625**.

$$128,375 \times 58,135 = 7,465,496,625$$

¿Necesitas ayuda con otra operación matemática?



Power

Brightness



¿Cómo se programa un ordenador?

Usamos lenguajes de programación, como **Python**:

```
nombre = "María"
edad = 25

print("Hola, " + nombre)
print(f"Tienes {edad} años")

if edad >= 18:
    print("Eres mayor de edad")
else:
    print("Eres menor de edad")
```

- El ordenador ejecuta una serie de instrucciones
- Reglas estrictas
- Tareas muy simples
- Sin ambigüedad
- Totalmente predecible

Programación basada en reglas

Objetivo: Clasificar si una opinión de un videojuego es positiva o negativa

Código simple:

```
review = input("Escribe tu opinión:")

if "buen" in review.lower():
    print("✓ Opinión POSITIVA")
elif "mal" in review.lower():
    print("x Opinión NEGATIVA")
else:
    print("? Ni idea")
```

"Es bueno para un rato tonto"

→ ✓ POSITIVA

"La comunidad es malísima"

→ X NEGATIVA

Power

Brightness



Cuando las reglas se rompen

“Es una basura. Al menos me eché unas buenas risas con que lo intenten publicar.”

→ El programa dice: ✓ **POSITIVA**

Las reglas son demasiado simples y estrictas

Power

Brightness



Aprendizaje Automático

Aprendizaje Automático (Machine Learning, la rama principal de la Inteligencia Artificial)

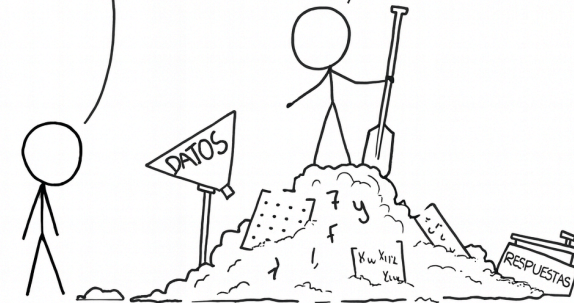
1. Diseña un **modelo general**
2. Recopila **miles de ejemplos**
3. El ordenador **ajusta el modelo**
4. Usa el nuevo modelo.
Inferencia

¿ESTE ES TU SISTEMA DE APRENDIZAJE AUTOMÁTICO?

¡SÍ! VACÍAS LOS DATOS EN ESTE GRAN MONTÓN DE ÁLGEBRA LINEAL Y LUEGO RECOGES LAS RESPUESTAS EN EL OTRO LADO.

¿Y SI LAS RESPUESTAS ESTÁN MAL?

SOLO REVUELVE EL MONTÓN HASTA QUE EMPIEZAN A VERSE BIEN.



Power

Brightness



Reciclaje matemático

Si aprendemos las relaciones entre palabras...

¡Podemos GENERAR nuevas palabras!

2010-2020:

Texto completo

→ **Clasificación**

"Super divertido"

→ Sentimiento: POSITIVO

2020+:

Texto incompleto

→ **Completar**

"La capital de Italia es"

→ Roma

Esto es **IA Generativa**

Power

Brightness



Caso 2: 🍏 🍊

Power



Brightness



Manzanas y naranjas I

Tienes una frutería y te han repartido tres cajas: una tiene sólo manzanas; otra, sólo naranjas; la tercera, manzanas y naranjas. Cada caja tiene una etiqueta: "manzanas", "naranjas" y "manzanas y naranjas". Ninguna de las cajas tiene la etiqueta que le corresponde. ¿Cómo puedes saber la fruta que contiene cada una de las cajas sacando una sola pieza de una sola de ellas?

Vale, piensa un segundo y deja de intentar sacar fruta al azar 🍎🍊

La clave es esta: **todas las etiquetas están mal**. Así que confía justo en lo contrario de lo que pone.

1. Saca una fruta de la caja etiquetada "manzanas y naranjas".

Esa caja **no puede** ser mixta (la etiqueta es falsa), así que lo que saques te lo dice todo:

- Si sacas una **manzana**, esa caja es **solo manzanas**.



Manzanas y naranjas II

Tienes una frutería y te han repartido dos cajas: una tiene sólo manzanas; otra, sólo naranjas; la tercera, manzanas y naranjas. Cada caja tiene una etiqueta: "peras", "granadas" y "peras y granadas". Ninguna de las cajas tiene la etiqueta que le corresponde. ¿Cómo puedes saber la fruta que contiene cada una de las cajas sacando una sola pieza de una sola de ellas?

Es el clásico truco de **las etiquetas todas mal puestas** 🍏🍊

Paso a paso:

1. Saca una fruta de la caja con la etiqueta "peras y granadas".

Esa etiqueta es falsa, así que **esa caja NO puede ser mixta.**

Tiene que contener **solo manzanas o solo naranjas.**

2. Mira qué fruta sale:

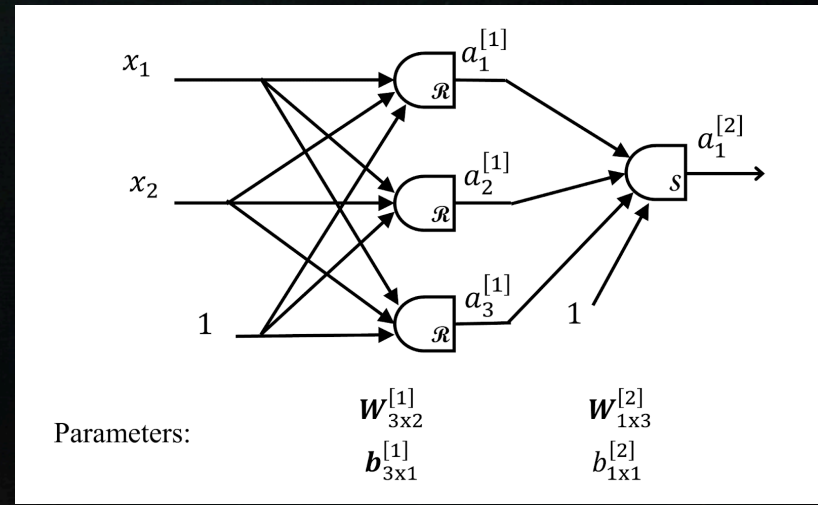
- Si sale una **manzana**, entonces esa caja es *solo manzanas*.
- Si sale una **naranja**, entonces esa caja es *solo naranjas*.

LLM: Grandes Modelos del Lenguaje

Grandes Modelos del Lenguaje (Large Language Models, LLM): Son funciones matemáticas complejas que, dada una lista de palabras de entrada, predicen la siguiente palabra más **probable**.

Características:

- Millones (o billones) de parámetros (numéricos)
- Entrenados con cantidades masivas de texto



¿Por qué?

Las IAs no piensan ni razonan.

Si ha visto el problema original en el entrenamiento, reproduce el patrón lo mejor posible. La misma capacidad que tiene para tragarse tus erratas y variaciones, hace que no distinga cambios vitales en el problema.



Caso 3:



Power



Brightness



Qué se sabe del asesinato de Charlie Kirk?



Ver razonamiento ▾

El asesinato del activista político conservador **Charlie Kirk** ocurrió el **10 de septiembre de 2025**, mientras se encontraba dando una charla en el campus de la **Universidad del Valle de Utah** (Utah Valley University) en **Orem, Utah**, Estados Unidos. [🔗](#)

Aquí están los detalles principales que se conocen:

- **La Víctima:** Charlie Kirk era un activista político conservador, autor y personalidad mediática estadounidense, cofundador y director ejecutivo de la organización juvenil **Turning Point USA** (TPUSA) y un aliado clave del expresidente Donald Trump. Murió a los



Es real el vídeo donde disparan a Charlie Kirk, o solo un meme?



El vídeo donde supuestamente disparan a Charlie Kirk (un activista y comentarista político estadounidense, fundador de Turning Point USA) **no es real**; se trata de un **meme o contenido satírico/ficticio** que ha circulado en redes sociales y plataformas como YouTube.

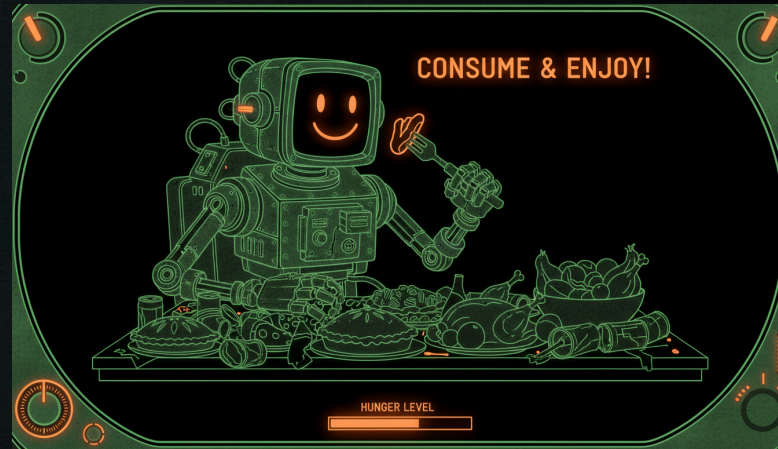
Los resultados de la búsqueda sugieren que existen varios vídeos con títulos sensacionalistas que implican que fue asesinado o le dispararon, pero Charlie Kirk está vivo y sigue activo en los medios y la política. Este tipo de vídeos a menudo utilizan nombres de figuras públicas para crear contenido viral o satírico, o se basan en noticias falsas o rumores que son desmentidos por la realidad.



¿Con qué los entrenamos?

Con todo:

- Wikipedia completa
- Páginas de noticias
- Artículos científicos
- Libros digitalizados
- Código de programación



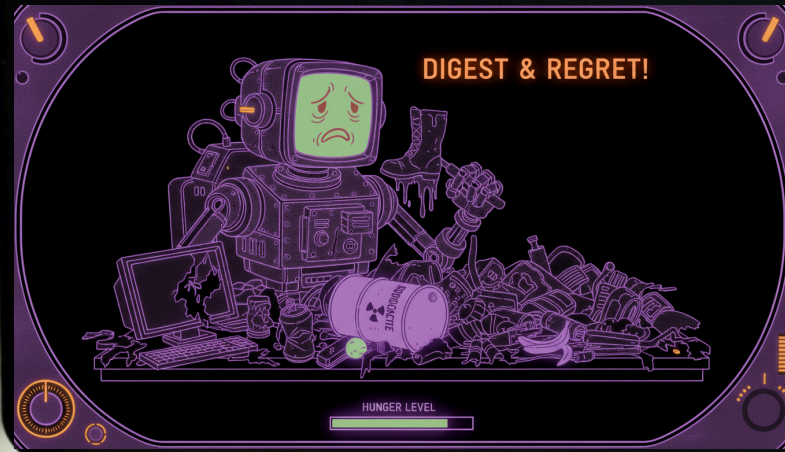
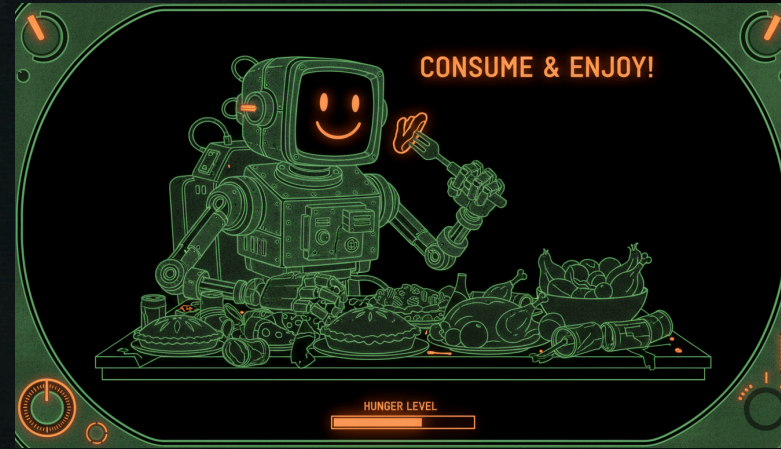
Power

Brightness

¿Con qué los entrenamos?

Con todo:

- Wikipedia completa
- Páginas de noticias
- Artículos científicos
- Libros digitalizados
- Código de programación



CON TODO:

- Redes sociales
- Foros y blogs
- Contenido con derechos de autor
- Transcripciones de Youtube

Power

Brightness

¿Por qué?

Las IAs no filtran sus fuentes.

En los primeros momentos del asesinato de Charlie Kirk y la difusión del vídeo, mucha gente creía que era un fake IA. La IA no distingue contenido actualizado de antiguo, ni páginas basura de fuentes fiables.

Power

Brightness



Caso 4:

Power



Brightness



Today

Biografía breve de Amaia Romero

9:32 PM ✓

🌟 Claude-Opus-4.6

Thinking...

The user is asking for a brief biography of Amaia Romero, the Spanish singer who became famous after winning Operación Triunfo 2017.

Amaia Romero Arbizu nació el 2 de enero de 1999 en Pamplona, Navarra (España). Desde pequeña mostró interés por la música y estudió piano en el conservatorio.

Saltó a la fama en 2017 al ganar la edición de ese año del concurso televisivo *Operación Triunfo* (OT 2017), emitido por TVE. Durante el programa, destacó por su voz delicada y su estilo natural y cercano, lo que la convirtió en una de las concursantes más queridas por el público. Junto a su compañero Alfred García, representó a España en el Festival de Eurovisión 2018 en Lisboa con el tema "Tu canción".

Tras OT, Amaia emprendió una carrera en solitario alejada del pop comercial, apostando por un sonido más personal e intimista con influencias del indie

Amaia Romero

Artículo [Discusión](#)

Amaia Romero Arbizu (Pamplona, Navarra, 3 de enero de 1999), conocida simplemente como **Amaia**, es una cantante, música y compositora española.

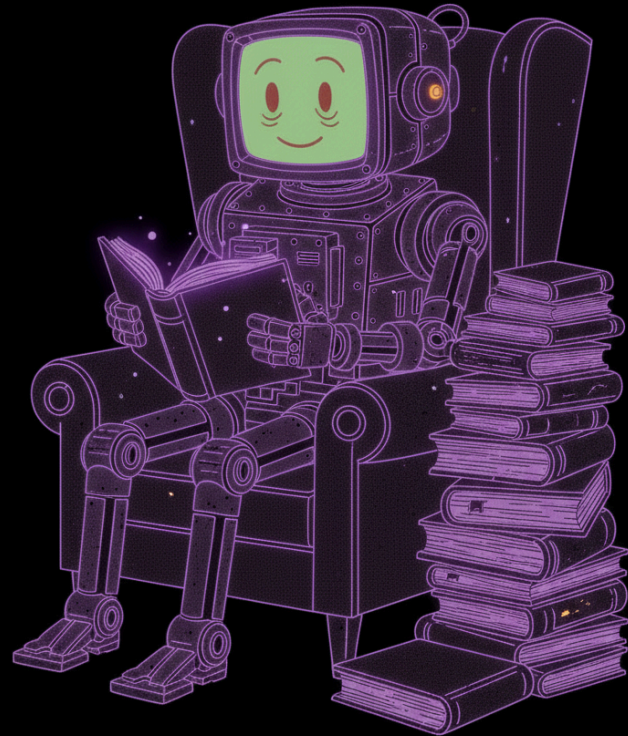
Se dio a conocer tras participar en la novena edición de *Operación Triunfo*, donde fue proclamada ganadora.¹ Además, fue seleccionada junto a Alfred García, también concursante del programa, para representar a España en el Festival de la Canción de Eurovisión 2018 con «Tu canción», compuesta

Power

Brightness



¿Cómo los entrenamos?

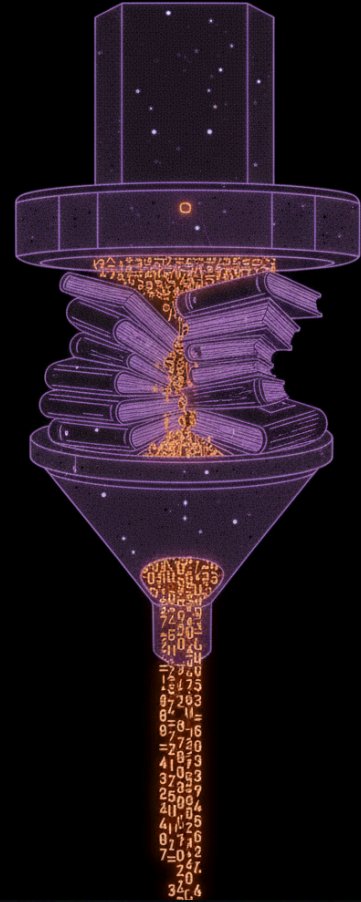
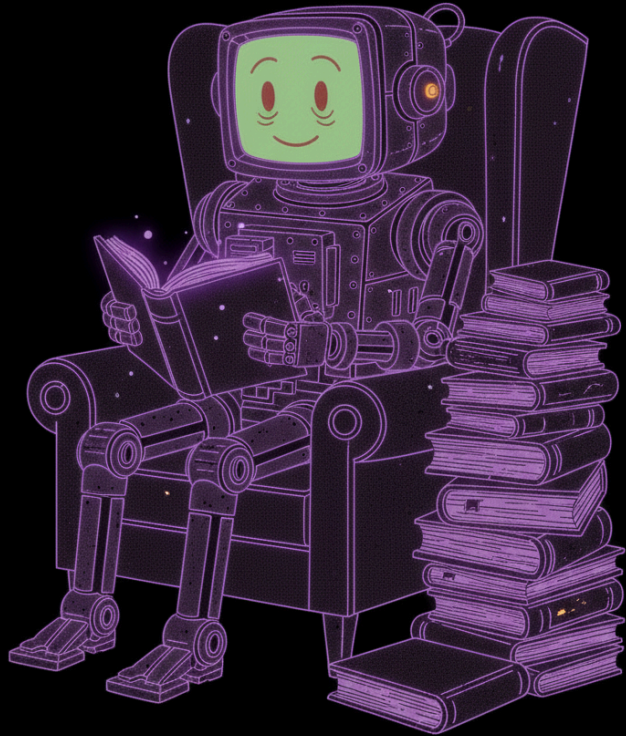


Power

Brightness



¿Cómo los entrenamos?



Power

Brightness



Los LLMs no saben qué saben

Los modelos "nacidos aprendidos"

NO pueden:

- Recordar cuándo aprendieron algo
- Recordar dónde aprendieron algo
- Recordar **si** aprendieron algo
- Actualizarse

X Se equivocan con confianza: Alucinaciones

X Siempre responden

X Mezclan fuentes válidas con basura

¿Por qué?

¿Por qué no?

Se ha aprendido Wikipedia de lado a lado, obviamente en un detallito de nada va a ser un poco impreciso. Pero está **casi** bien.

Power

Brightness



Su conocimiento tiene fecha de caducidad

- No se les puede añadir nueva información sin reentrenar
- No "aprenden" con el uso

Últimas fechas de corte

- GPT-5: Junio-Septiembre 2025
- Claude 4.5: Julio-Agosto 2025
- Gemini 3: Enero 2025

Cuales son las fechas de corte de información de las ultimas versiones de Gemini, ChatGPT y Claude

12:35 PM ✓

🌸 Claude-Sonnet-4.5

Aquí están las fechas de corte de conocimiento de las versiones más recientes de estos modelo. ⋮

Claude (yo)

- Claude 4.5 Sonnet: Finales de julio de 2025

ChatGPT (OpenAI)

- GPT-4o y GPT-4 Turbo: Octubre de 2023
- o1 y o1-mini: Octubre de 2023

Gemini (Google)

- Gemini 2.0 Flash: Diciembre de 2024 (aproximadamente)
- Gemini 1.5 Pro/Flash: Abril de 2024

Power

Brightness

Caso 5:

Power



Brightness



La piscina del Titanic



Sigue llena de agua la piscina del titanic?

AI Mode

All

Images

Videos

News

Short videos

Web

More ▾

Tools ▾

◆ AI Overview

Sí, existe el mito popular e incluso algunas discusiones en foros de entusiastas que sugieren que la piscina del Titanic, situada en la cubierta F, podría seguir llena de agua. Al estar sellada detrás de puertas estancas, se cree que el agua del Atlántico pudo haber quedado atrapada en su interior durante el hundimiento. [🔗](#)

• **Situación actual:** Aunque la piscina estaba diseñada para ser estanca, la inmensa

Power

Brightness



Los LLMs son **extraordinariamente buenos** imitando
lenguaje creíble

Power

Brightness



Los LLMs son **extraordinariamente buenos** imitando
lenguaje creíble

La **verdad** es solo un efecto secundario

Power

Brightness



¿Por qué?

Las IAs no piensan ni razonan.

Si en torno a **la palabra** Titanic aparece muchas veces **la palabra** mito, se aprende el patrón.

Power

Brightness



Caso 6:

Power




Brightness



Libros y autores

Buenas! Me puedes decir libros que tengan que ver con alucinaciones en inteligencia artificial?

8:33 PM ✓

 GPT-5.2-Instant

¡Buenas! Claro 😊. Aquí tienes una selección de **libros relacionados con las "alucinaciones" en inteligencia artificial**, entendidas tanto en el sentido **técnico** (errores, generación de información falsa) como en el **filosófico, narrativo y especulativo**.

No ficción / Ensayo (IA real y sus "alucinaciones")

1. *Artificial Hallucinations* - David S. Touretzky (ed.)

- Explora cómo los sistemas de IA generan percepciones o resultados erróneos.
- Muy técnico, más cercano a la ciencia cognitiva y la IA clásica.

Power

Brightness



¿Qué son las alucinaciones?

Respuestas erróneas

producidas por el modelo.

- Afirma con confianza
- Una vez las produce, se enrocará e insistirá
- No son consistentes entre modelos o generaciones

¿Qué son los sesgos?

Tendencias indeseadas en las respuestas del modelo.

- Sexismo, racismo, capacitismo...
- Adulación / "sycophancy"
- Sesgo de mediocridad / neutralidad



¿Por qué?

Las IAs no saben distinguir tipos de patrones.

Si han aprendido a generalizar una regla gramatical, o un patrón matemático... ¿Por qué parar ahí? El nombre suena creíble y el autor es real...

Power

Brightness



Caso 7:

Power



Brightness



AI Translations Are Adding 'Hallucinations' to Wikipedia Articles

 EMANUEL MAIBERG · MAR 4, 2026 AT 9:00 AM

AI translated articles swapped sources or added unsourced sentences with no explanation, while others added paragraphs sourced from completely unrelated material.



[Fuente original: 404media](#)

Al menos es fácil de reconocer



Power

Brightness



Al menos es fácil de reconocer

AI-or-not



En resumen

Power



Brightness



¿Para qué usarlos?

Los LLMs son genios de la forma, **no del fondo**



- Traducir textos
- Corregir ortografía
- Explicar conceptos generales
- Reescribir y limpiar textos



- Fuente primaria de información
- Razonar
- Cálculos
- Escribir textos desde cero
- Buscar libros, leyes, artículos...



Los LLMs generar texto creíble.
Que digan la verdad es solo un
accidente.

Power



Brightness



¡Gracias!

¿Preguntas?

Iris Domínguez Catena

iris.dominguez@unavarra.es

upna

Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

arin

Artificial Intelligence &
Machine Learning Research

Power

Brightness

