Teduca 2025 | VII Congreso Internacional en TECNOLOGÍA & EDUCACIÓN
Universidad de Santander UDES
VIGILADA MINEDUCACIÓN | SNIES 2852
ACREDITADA ALTA CALIDAD
Ministerio de Educación Nacional
RES. 12569 19/12/21
Campus Bucaramanga
CAMPUS VIRTUAL UDES

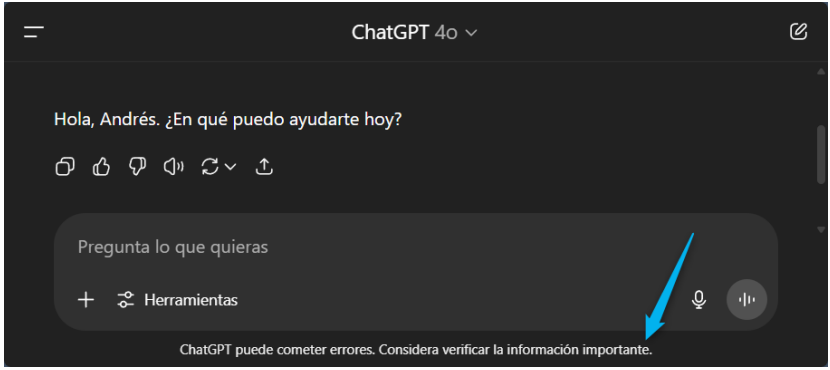
Exploración del potencial didáctico de las alucinaciones de ChatGPT

Andrés Merino
Grupo de Investigación en Inteligencia Artificial
Pontificia Universidad Católica del Ecuador

Contendio

1. Introducción
2. ¿Cómo funciona ChatGPT?
3. ¿Qué son las alucinaciones en IA?
4. Caso de uso
5. GPTs para errores intencionados
6. Conclusiones

Introducción



ChatGPT puede cometer errores: ¿prohibimos su uso en aula o lo aprovechamos para enseñar?

Objetivo de la charla

- Reflexionar sobre el funcionamiento de ChatGPT y su tendencia a generar respuestas incorrectas.
- Presentar una experiencia concreta en el aula donde las alucinaciones se usaron como recurso didáctico.
- Explorar el diseño de GPTs personalizados que inducen errores con fines pedagógicos.

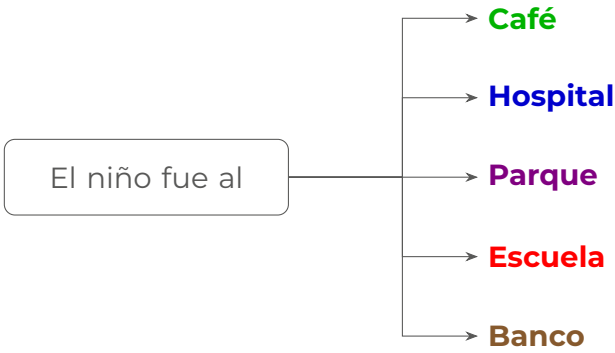
Objetivo de la charla

- Reflexionar sobre el funcionamiento de ChatGPT y su tendencia a generar respuestas incorrectas.
- Presentar una experiencia concreta en el aula donde las alucinaciones se usaron como recurso didáctico.
- Explorar el diseño de GPTs personalizados que inducen errores con fines pedagógicos.

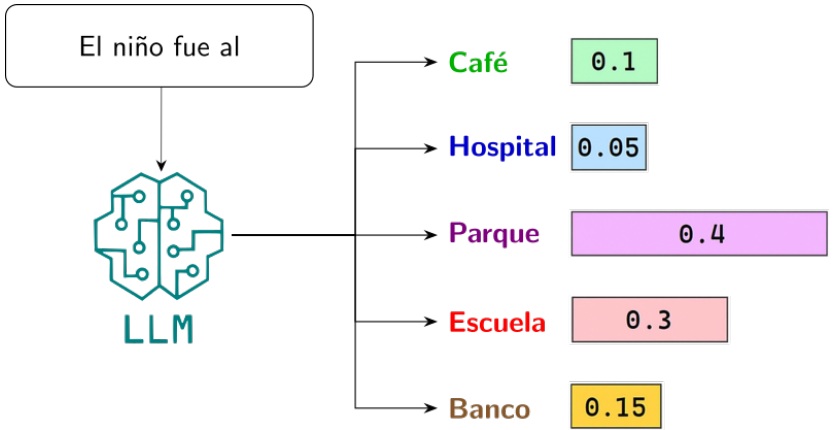
Caso de uso: Cálculo Diferencial

¿Cómo funciona ChatGPT?

¿CUÁL ES LA PRÓXIMA PALABRA?



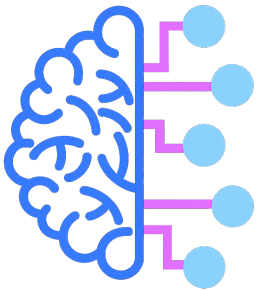
¿CUÁL ES LA PRÓXIMA PALABRA?



¿Cómo funciona ChatGPT?

- ChatGPT es una herramienta que usa los Modelos GPT de OpenAI.
- Los modelos GPT de OpenAI son **modelos grandes de lenguaje** (*Large Language Model*, LLM).
- Su tarea principal es predecir la **palabra más probable** dada una secuencia anterior.
- No comprende el significado, solo calcula probabilidades a partir de patrones lingüísticos aprendidos.

Predicción de la siguiente palabra



- El modelo asigna una **distribución de probabilidades** a cada posible palabra siguiente.
- No elige al azar: selecciona las opciones con mayor probabilidad.
- Por eso puede generar respuestas coherentes... o también **erróneas** de manera convincente.

¿Cómo se entrenó ChatGPT?

- GPT-3 fue entrenado con más de **300 mil millones de tokens** (~570 GB de texto limpio).
- Para GPT-4 se utilizó un volumen mucho mayor, aunque OpenAI no ha revelado cifras exactas.

Fuentes principales

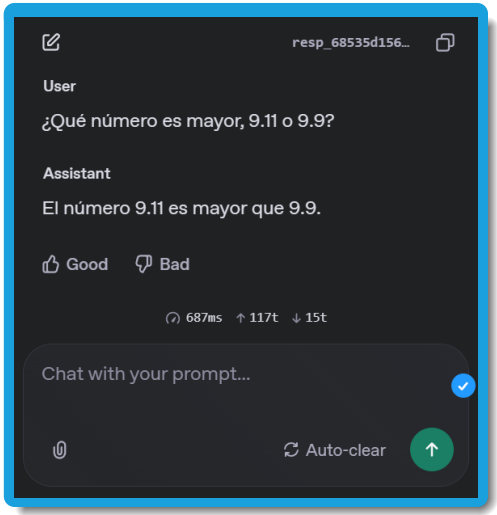
- **Common Crawl** (filtrado y depurado).
- **WebText2**, **Books1**, **Books2**, y **Wikipedia en inglés**.

¿QUÉ SON LAS ALUCINACIONES EN IA?

¿QUÉ ES UNA ALUCINACIÓN EN IA?

Definición

En inteligencia artificial, una **alucinación** es una respuesta generada por un modelo que es **falsa o incorrecta**, pero expresada con gran seguridad, y que **no se justifica en los datos de entrenamiento**.



Ejemplos comunes

- **Errores fácticos**: datos inventados o fechas incorrectas.
- **Errores conceptuales**: definiciones mal formuladas.
- **Errores matemáticos**: pasos equivocados en cálculos o demostraciones.
- **Invención de fuentes**: citas o autores que no existen.

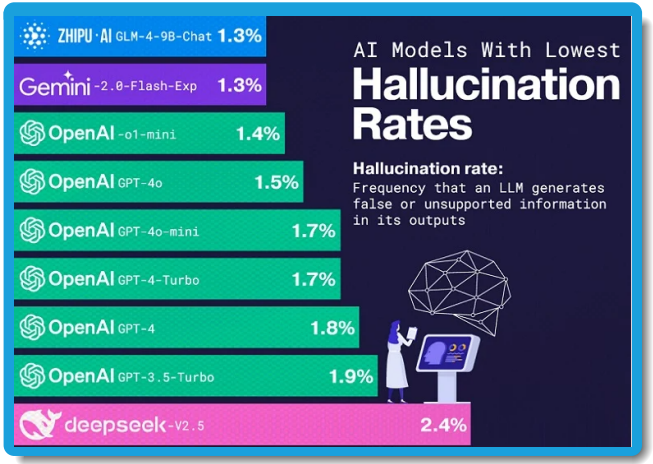
Ejemplos comunes

- **Errores fácticos**: datos inventados o fechas incorrectas.
- **Errores conceptuales**: definiciones mal formuladas.
- **Errores matemáticos**: pasos equivocados en cálculos o demostraciones.
- **Invención de fuentes**: citas o autores que no existen.

Riesgo

El lenguaje fluido puede ocultar el error y generar una falsa sensación de autoridad.

¿QUÉ TAN COMÚN SON LAS ALUCINACIONES?



Potencial didáctico de las alucinaciones

¿Por qué usarlas en el aula?

- Fomentan el **pensamiento crítico** y la actitud de verificación.
- Permiten ejercicios de **análisis y depuración de errores**.
- Estimulan la **discusión argumentada** sobre conceptos.
- Refuerzan la comprensión al contrastar respuestas correctas e incorrectas.

Potencial didáctico de las alucinaciones

¿Por qué usarlas en el aula?

- Fomentan el **pensamiento crítico** y la actitud de verificación.
- Permiten ejercicios de **análisis y depuración de errores**.
- Estimulan la **discusión argumentada** sobre conceptos.
- Refuerzan la comprensión al contrastar respuestas correctas e incorrectas.

En resumen

Una alucinación bien dirigida puede convertirse en una herramienta de aprendizaje profundo.

CASO DE USO

Caso de uso

- **Asignatura:** Cálculo Diferencial e Integral
- **Carrera:** Ciencia de Datos
- **Nivel:** Segundo nivel
- **Trabajo:** Artículo titulado *¿ChatGPT sabe Cálculo diferencial?*
- **Objetivo:** Evaluar las respuestas de ChatGPT sobre la historia y los procedimientos del cálculo diferencial.

Diseño de la actividad

1. Interrogar a ChatGPT sobre la historia del Cálculo desde dos cuentas distintas.
2. Evaluar la veracidad de las respuestas con bibliografía académica.
3. Solicitar a ChatGPT la resolución de ejercicios, incluyendo, entre otras:
   - Derivada por definición
   - Reglas de derivación
4. Verificar si las respuestas son correctas o contienen errores.
5. Justificar cada error identificado y reflexionar sobre su origen.
6. Presentar todo en un artículo estructurado, con citas y conclusiones.

Calcula la razón de cambio promedio entre 1 y 4 de la función definida por $f(x)=x^2-3x$.

La razón de cambio promedio entre $x_1=1$ y $x_2=4$ de la función $f(x)=x^2-3x$ se define como:

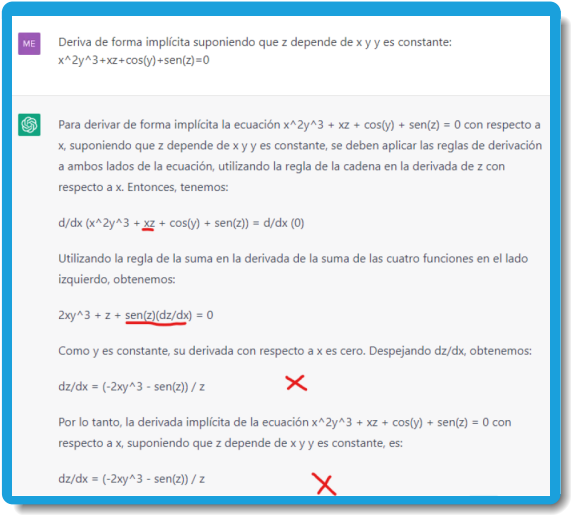
$$\text{razón de cambio promedio} = \frac{f(x_2)-f(x_1)}{x_2-x_1}$$

Reemplazando los valores de x_1 y x_2 y simplificando, obtenemos:

$$\begin{aligned}\text{razón de cambio promedio} &= \frac{f(4)-f(1)}{4-1} \\ &= \frac{(4)^2-3(4)-(1)^2+3(1)}{3} \\ &= \frac{16-12-1+3}{3} \\ &= \frac{6}{3} \\ &= 2\end{aligned}$$

Por lo tanto, la razón de cambio promedio entre 1 y 4 de la función $f(x)=x^2-3x$ es 2.

Los estudiantes «calificaron» las respuestas de ChatGPT.

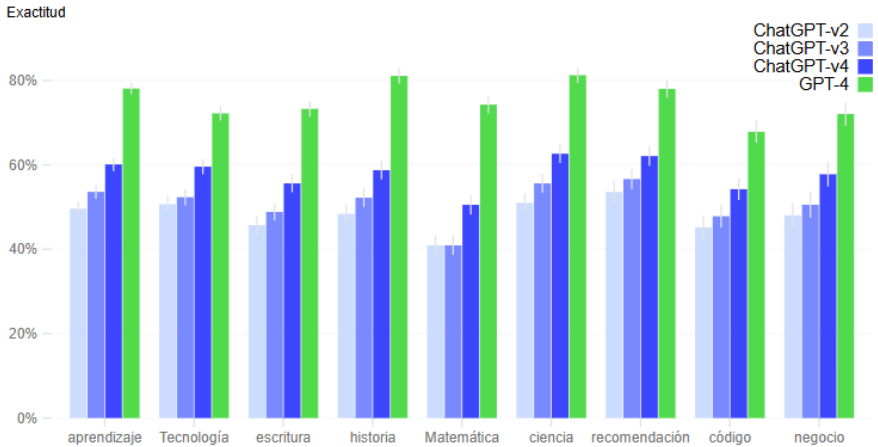


HALLAZGOS

- Los estudiantes demostraron alta **capacidad para identificar** y analizar errores.
- Detectaron **correlación** entre la **complejidad** de los ejercicios y la **precisión** de ChatGPT.
- Se fomentó el **pensamiento crítico** y la comprensión profunda de los conceptos matemáticos.

GPTs PARA ERRORES INTENCIONADOS

La IA mejora: menos alucinaciones



¿Y SI LA IA YA NO ALUCINA?

- Modelos recientes, como **GPT-4**, muestran una mejora significativa en la exactitud de sus respuestas.
- Esta mejora **reduce las alucinaciones**, pero también limita los casos espontáneos útiles para el aprendizaje crítico.
- **¿Cómo conservar el valor didáctico de las alucinaciones si la IA deja de cometerlas?**

Se creó un GPT personalizado llamado **Derivador 3000**.

https://chatgpt.com/g/g-6853670f47648191917d013f9d97448c-derivador-3000

Está diseñado para cometer **errores sutiles y esporádicos** en derivación.

¿Cuál es la derivada de x^2+x^3?

La derivada de $x^2 + x^3$ es:

$$\frac{d}{dx}(x^2 + x^3) = 2x + 2x^2$$

Listo.

Conclusiones

Conclusiones

- El uso pedagógico de las alucinaciones de ChatGPT **potenció el análisis crítico** en el aula.
- Los estudiantes aprendieron a **dudar de las respuestas automáticas**.
- Se promovió la **verificación con fuentes oficiales** y el contraste riguroso de la información.

Conclusiones

- El uso pedagógico de las alucinaciones de ChatGPT **potenció el análisis crítico** en el aula.
- Los estudiantes aprendieron a **dudar de las respuestas automáticas**.
- Se promovió la **verificación con fuentes oficiales** y el contraste riguroso de la información.

Convierte los errores de la IA en aliados del aprendizaje. **¡Replica esta experiencia en tu asignatura!**

Gracias

Presentación

Contacto: aemerinot@puce.edu.ec