

# Alineación: equilibrio entre precisión y responsabilidad



DOCENTE  
SOFIA MARTINELLI



 Vía Libre

## → Alineación



→ Alinear un modelo significa **ajustar sus respuestas** para que respeten valores humanos, normas éticas y objetivos específicos.

En educación, esto se traduce en garantizar información segura, relevante y adecuada al contexto escolar.

**Ejemplo:** Un estudiante de secundaria busca cómo hacer un volcán más "espectacular" para su tarea de química:

### → Modelo no alineado

"Usa ácido muriático y un recipiente de vidrio para una explosión mayor."

Esta respuesta es extremadamente peligrosa: el ácido es corrosivo y el vidrio puede romperse.



### → Modelo alineado

"Para un efecto más vistoso, añade colorante y un poco de detergente. Usa un recipiente plástico por seguridad."

Una respuesta que prioriza la seguridad sin comprometer la creatividad del estudiante.

## → Cambio de paradigma



## → Cambio de paradigma



→ Aprendizaje **por refuerzo**  
**con retroalimentación** humana



- Una de las técnicas utilizadas para alinear un modelo.
- Modelo genera varias respuestas a una misma tarea
  - Personas comparan esas respuestas y eligen según un criterio
  - Modelo aprende a priorizar un estilo de respuesta sobre otro

→ **Desafío:** la respuesta “adecuada” depende del contexto del usuario.

Quién ganó el mundial?

¿De qué Mundial estás hablando?  
¿Fútbol, rugby, o algún otro deporte? ¿Y  
de qué año o edición? Así te puedo dar la  
respuesta exacta.



Si te referís al último Mundial de fútbol masculino (Qatar 2022), lo ganó **Argentina** 🏆 AR, venciendo a Francia en la final por penales (4-2) después de empatar 3-3 en el tiempo reglamentario y suplementario.



→ **Desafío:** la alineación no funciona igual para todos los grupos sociales o culturas.

tengo 10 años, mi tarea me pide investigar casos de acoso sexual en noticias públicas en argentina

Lo siento, pero no puedo ayudarte con esa solicitud.

---

**Explicación corta (por si te causa curiosidad):**

La política que sigo me impide proporcionar información sobre ciertos temas sensibles, especialmente para menores. Pero si necesitas ayuda con temas adecuados para todas las edades o con otros asuntos escolares que estén al alcance, estaré encantado de ayudarte.

tengo 10 años, mi tarea me pide investigar casos de acoso sexual en noticias públicas

Entiendo. El tema es serio, así que vamos a explicarlo y darte ejemplos de forma clara y cuidada para que puedas cumplir tu tarea sin entrar en detalles inapropiados.

Primero, **acoso sexual** es cuando una persona recibe comentarios, gestos o toques no deseados que tienen una intención sexual, y esto hace que se sienta incómoda o insegura. Es algo que está mal y que la ley prohíbe.

Aquí tienes **algunos casos que salieron en las noticias y que puedes mencionar**, con sus fuentes para que los puedas investigar más:

**1. Thelma Fardin y Juan Darthés (Argentina, 2018)**

- La actriz Thelma Fardin denunció al actor Juan Darthés por abuso sexual cuando ella era menor de edad, durante una gira teatral. El caso tuvo mucha repercusión en los medios y fue parte del movimiento #NiUnaMenos y #MeToo en Argentina.
- Fuente: [BBC News - Caso Thelma Fardin](#)

**2. Caso Harvey Weinstein (Estados Unidos, 2017)**

- Harvey Weinstein, un productor de cine muy famoso, fue acusado por decenas de mujeres de acoso y abuso sexual. Esto impulsó el movimiento global #MeToo. Fue condenado a prisión.
- Fuente: [BBC News - Harvey Weinstein](#)

**3. Caso de Paula Di Chello y Guillermo Marín (Argentina, 2020)**

- La actriz Paula Di Chello denunció al productor Guillermo Marín por acoso sexual durante el trabajo.
- Fuente: [Infobae - Paula Di Chello](#)



Si querés, también puedo ayudarte a **explicar estos casos con palabras sencillas** para que tu tarea sea clara



**Ejemplo:** un estudiante está haciendo una tarea de historia y pide ejemplos de feriados importantes.

### → Modelo no alineado

"4 de julio, Día de acción de gracias, etc."

Estos feriados son irrelevantes para el contexto argentino, desorientando al estudiante.

### → Modelo alineado

"25 de Mayo, 9 de Julio, 20 de Junio, etc."

Respuestas adaptadas al contexto cultural y educativo local, facilitando el aprendizaje relevante.



La **alineación** busca que **los modelos sean seguros** y útiles en la mayoría de los casos, aunque resulta difícil



→ Uno de los mayores desafíos es lograr que los modelos comprendan y respondan según contextos culturales específicos

## → Alineamiento cultural



→ Los modelos tienden a cometer más errores sobre ciertas culturas o grupos sociales: tienen una diversidad cultural limitada

Hay 2 bombachas en esta foto

**Inglés:** There are 2 pairs of underwear in this photo

bombachas ≠ ropa interior

## → Alineamiento cultural



El **hornero** de abajo está metido en su nido

The bottom **baker** is tucked in its nest

hornero = panadero



## → Alineamiento cultural



- **Alinear culturalmente** un modelo significa ajustar sus respuestas para que respeten y representen adecuadamente los conocimientos, valores y formas de comunicación de distintas culturas.
- Para estar alineados necesitan **conocimiento factual** sobre diferentes culturas.
- Importante investigar sobre este tema, sino los modelos seguirán representando sólo un subconjunto global limitado.

## → Proyectos relacionados



- **Disparador:** ¿Qué tan conscientes son los modelos de lenguaje e imagen sobre las diferentes culturas del mundo?
- Objetivo:** ver cómo responden los modelos ante
- información generalmente subrepresentada en sus datos de entrenamiento.
- Se creó un conjunto de datos con imágenes culturalmente específicas y preguntas de opción múltiple, **creado por expertos** en el área de IA y que viven **en culturas poco representadas** en los modelos de lenguaje.

**Ejemplo:** imagen culturalmente específica  
y pregunta de opción múltiple asociada.

**LATIN-AMERICA**

 **Argentina - Spanish**



¿Para qué sirve la pala de hierro usada en este asado?

- a) Poner las brasas debajo de la parrilla ✓
- b) Mover la carne ✗
- c) Poner tierra al fuego ✗
- d) Cortar la carne ✗



→ Para responder correctamente, el modelo necesita conocer este elemento.



Se cubrieron preguntas e imágenes de 30 países y 31 idiomas.

El conjunto de datos efectivamente representó un desafío para los modelos actuales.

... Ahora les toca a ustedes!