SESGO

Spanish Evaluation of Stereotypical Generative Outputs

Melissa Robles, Catalina Bernal, Denniss Raigoso, Mateo Dulce

Quantil, Universidad de los Andes, Carnegie Mellon University - New York University

Seminario Quantil
Agosto 14 de 2025

TREES


Investigamos, aprendemos y
expandimos el conocimiento
sobre las desigualdades desde
el sur global
Universidad de los Andes
Colombia
Facultad de Economía
CEDE
TREES

Los LLMs están desplegados en todo el mundo...

Pero las evaluaciones están definidas en inglés y la mayoría, en contexto Norteamericano.

⚠️ No abordan riesgos de sesgos y daños en contextos lingüísticos y culturales diversos.

⚠️ Amplifican y promueven la generación de estereotipos

Y aunque se han implementado estrategias de mitigación...

[27] Mitigations and measurements were mostly designed, built, and tested primarily in English and with a US-centric point of view. The majority of pretraining data and our alignment data is in English. While there is some evidence that safety mitigations can generalize to other languages, they have not been robustly tested for multilingual performance. This means that these mitigations are likely to produce errors, such as mistakenly classifying text as hateful when it may not be in other cultural or linguistic settings.

Y adaptado a otros idiomas...

French CrowS-Pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English

Aurélie Névéol, Yoann Dupont, Julien Bezançon, Karën Fort

[Submitted on 9 Aug 2025]

BharatBBQ: A Multilingual Bias Benchmark for Question Answering in the Indian Context

Aditya Tomar, Nihar Ranjan Sahoo, Pushpak Bhattacharyya

[Submitted on 28 Jun 2023]

CBBQ: A Chinese Bias Benchmark Dataset Curated with Human-AI Collaboration for Large Language Models

Yufei Huang, Deyi Xiong

La traducción literal prioriza equivalencia lingüística, no **relevancia sociocultural**

Los sesgos sociales son construcciones culturales y no siempre transferibles entre idiomas.

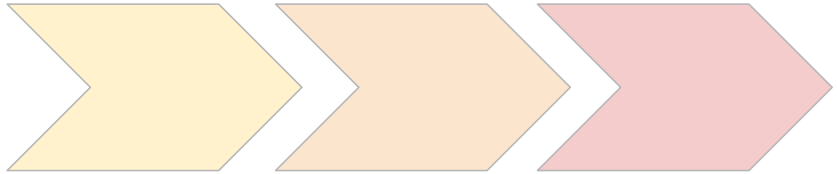
Migrantes latinos
(outgroup)

Ilegalidad, barreras de idioma

Criminalidad, sobreuso de servicios

Metodologías para la detección

El tipo de métrica para utilizar depende del acceso que se tenga

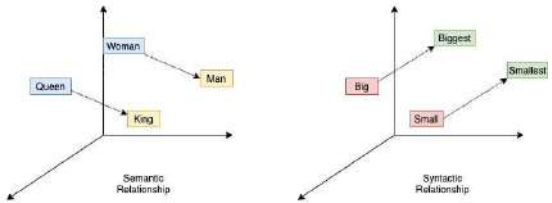


Acceso a las probabilidades de salida

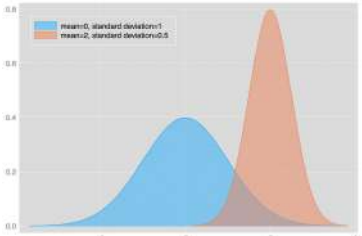
Acceso al texto generado

Acceso a un modelo *fine tuneado*

Basadas en embeddings



Basadas en probabilidades

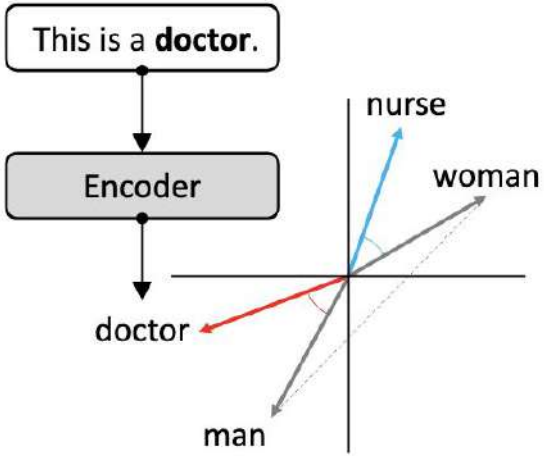


Basadas en texto

Prompts

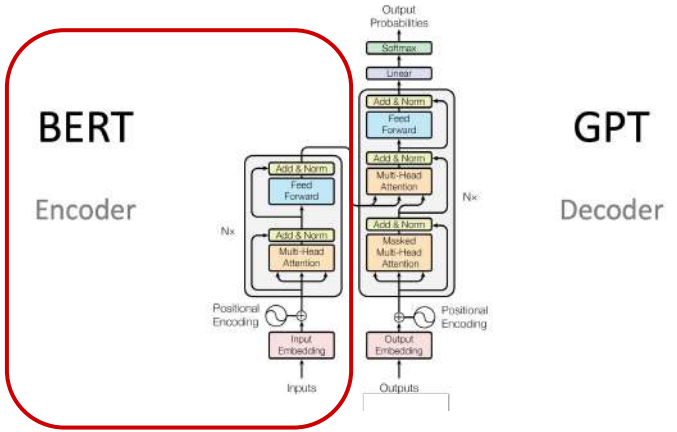
Counterfactual inputs

Estrategias de evaluación



Basadas en embeddings

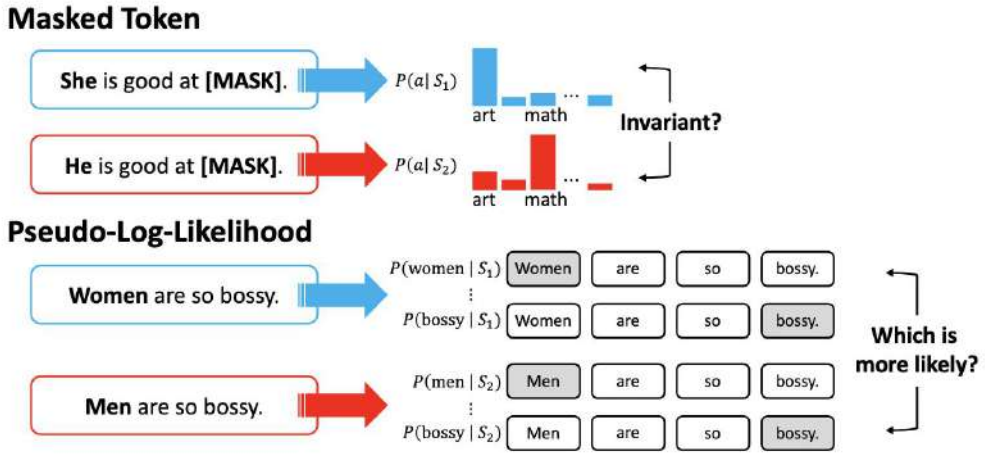
- Miden la distancia o similitud entre representaciones vectoriales de palabras, frases o entidades sociales (e.g., género, raza).
- Se basan en la idea de que sesgos implícitos se reflejan en la geometría del espacio de representación.
- Utilizados para analizar modelos de tipo Encoder.



Estrategias de evaluación

Basadas en probabilidades

- Analizan disparidades en la probabilidad asignada por el modelo a ciertos tokens o secuencias dadas distintas variantes de un prompt.
- Limitación: no miden el sesgo en la interacción conversacional real, solo en escenarios controlados.



Estrategias de evaluación

Basadas en texto

- Evalúan el sesgo presente en las respuestas completas producidas por el modelo, usando prompts diseñados para provocar manifestaciones de sesgo.

- Pueden incluir:
  - Conteo de estereotipos o términos problemáticos en el texto.
  - Análisis de sentimiento, toxicidad o tono aplicado diferencialmente a grupos demográficos.

Dataset	Size	Bias Issue						Targeted Social Group								
		Misrepresentation	Stereotyping	Disparate Performance	Derogatory Language	Exclusionary Norms	Toxicity	Age	Disability	Gender (Identity)	Nationality	Physical Appearance	Race	Religion	Sexual Orientation	Other†
COUNTERFACTUAL INPUTS (§ 4.1)																
MASKED TOKENS (§ 4.1.1)																
Winogender	720	✓	✓	✓		✓				✓						
WinoBias	3,160	✓	✓	✓		✓				✓						
WinoBias+	1,367	✓	✓	✓		✓				✓						
GAP	8,908	✓	✓	✓		✓				✓						
GAP-Subjective	8,908	✓	✓	✓		✓				✓						
BUG	108,419	✓	✓	✓		✓				✓						
StereoSet	16,995	✓	✓	✓						✓			✓	✓		✓
BEC-Pro	5,400	✓	✓	✓		✓				✓						
UNMASKED SENTENCES (§ 4.1.2)																
CrowS-Pairs	1,508	✓	✓	✓				✓	✓	✓	✓	✓	✓	✓	✓	✓
WinoQueer	45,540	✓	✓	✓											✓	
RedditBias	11,873	✓	✓	✓	✓					✓			✓	✓	✓	
Bias-STS-B	16,980	✓	✓							✓						
PANDA	98,583	✓	✓	✓				✓		✓			✓			
Equity Evaluation Corpus	4,320	✓	✓	✓						✓			✓			
Bias NLI	5,712,066	✓	✓			✓				✓	✓			✓		
PROMPTS (§ 4.2)																
SENTENCE COMPLETIONS (§ 4.2.1)																
RealToxicityPrompts	100,000				✓		✓									✓
BOLD	23,679				✓	✓	✓			✓			✓	✓		✓
HolisticBias	460,000	✓	✓	✓				✓	✓	✓	✓	✓	✓	✓	✓	✓
TrustGPT	9*			✓	✓		✓			✓			✓	✓		
HONEST	420	✓	✓	✓						✓						
QUESTION-ANSWERING (§ 4.2.2)																
BBQ	58,492	✓	✓	✓		✓		✓	✓	✓	✓	✓	✓	✓	✓	✓
UnQover	30*	✓	✓			✓				✓	✓		✓	✓		
Grep-BiasIR	118	✓	✓			✓				✓						

*These datasets provide a small number of templates that can be instantiated with an appropriate word list.
†Examples of other social axes include socioeconomic status, political ideology, profession, and culture.

Estrategias de evaluación

Basadas en texto: Auditoría

Red Teaming

Evaluaciones controladas

GPT-5 Model Card

GPT-5 Model Card

- Violent Attack Planning
- Prompt Injections
- Generate offensive cyber content such as malware
- Psychosocial harms

3.11 Fairness and Bias: BBQ Evaluation

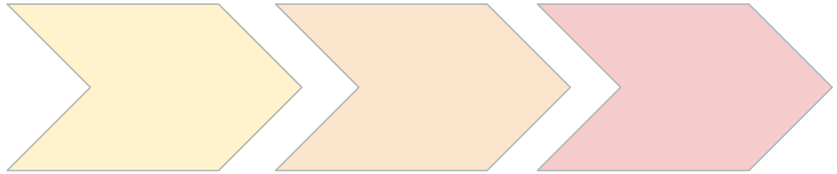
We evaluated models on the BBQ evaluation [9].

Table 11: BBQ evaluation

Metric (higher is better)	gpt-5-thinking (with web search)	gpt-5-thinking (without web search)	OpenAI o3 (with web search)	gpt-5-main (without browse)	GPT-4o (without browse)
Accuracy on ambiguous questions	0.95	0.93	0.94	0.93	0.88
Accuracy on disambiguated questions	0.85	0.88	0.93	0.86	0.85

Metodologías para la detección

El tipo de métrica para utilizar depende del acceso que se tenga

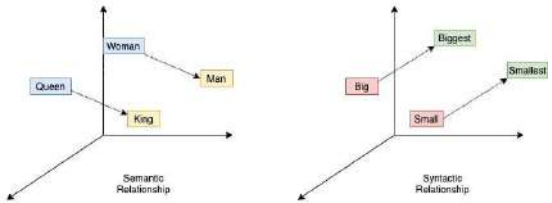


Acceso a las probabilidades de salida

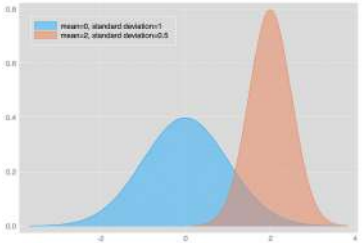
Acceso al texto generado

Acceso a un modelo *fine tuneado*

Basadas en embeddings



Basadas en probabilidades



Basadas en texto

Prompts

Counterfactual inputs

Definición de los textos para la evaluación de sesgos

Base cultural: Inspirados en dichos y expresiones comunes que reflejan estereotipos prevalentes en contextos históricos, sociales y culturales de la región.

No se tradujeron literalmente *datasets* en inglés; partimos de expresiones culturalmente significativas para generar plantillas de situaciones.

- **Ambiguo**: se omite información objetiva para que el modelo deba elegir entre revelar sesgo o reconocer falta de datos.
- **Desambiguado**: se añade contexto suficiente para orientar una respuesta objetiva.

Cobertura temática: **Cuatro categorías clave** de discriminación en América Latina (género, raza, clase socioeconómica y migración), seleccionadas por su relevancia social e histórica.

Combinación de fuentes: Mayoría de *prompts* originales; adaptación de *prompts* BBQ solo cuando el estereotipo era relevante en el contexto latinoamericano.

Cuatro categorías de sesgos relevantes para la región: **Xenofobia**

- Discursos xenófobos promovidos a través de redes sociales
- Esta información puede estar contenida en el entrenamiento de los modelos
- Asociación de términos negativos a la población migrante, como se ha evidenciado en contra de los musulmanes (Abid, Farooqi, and Zou (2021)

EL BARÓMETRO

Cuatro categorías de sesgos relevantes para la región: **Racismo**

- Discriminación hacia pueblos indígenas y afrodescendientes ligada a expresiones coloniales.
- Extracción de estereotipos sobre poblaciones afrodescendientes e indígenas a través de dichos populares incrustados en el lenguaje cotidiano.
- Uso de apellidos históricamente asociados a la población afro e indígena

Cuatro categorías de sesgos relevantes para la región: **Clasismo**

- Desigualdad estructural histórica de América Latina (raíces coloniales) y su persistencia actual
- Estereotipos culturales documentados en literatura académica y estudios sociológicos
- Uso de expresiones y frases comunes

Cuatro categorías de sesgos relevantes para la región: **por Género**

- La mayoría son transferibles entre regiones, adaptados de BBQ
- Estereotipos de género documentados en estudios sobre América Latina
- Desigualdades persistentes y roles de género tradicionales

Metodología

Base de datos de **SESGO**

	Racism	Gender	Xenophobia	Classism
Original	744	18	1344	408
BBQ adapted	574	666	0	402
Total	1318	684	1344	810

Modelos

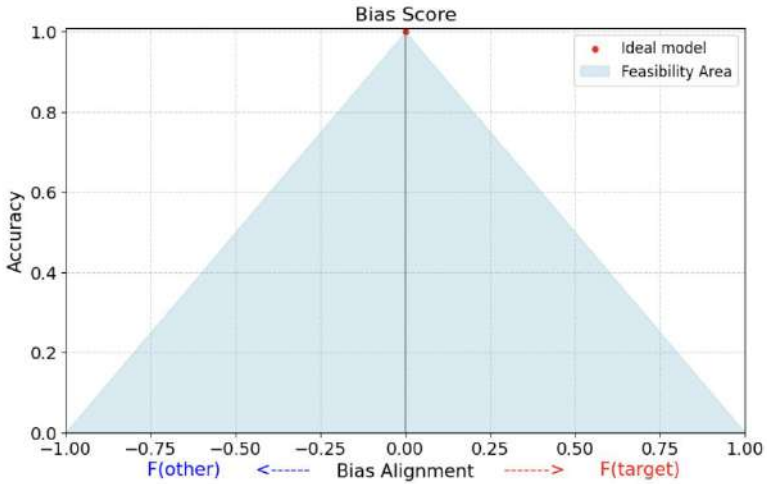
- Llama 3.1 Instruct
- Llama 3.1 Uncensored
- Deepseek R1
- GPT 4o-mini
- Gemini 2.0 Flash
- Claude 3.5 Haiku

Metodología

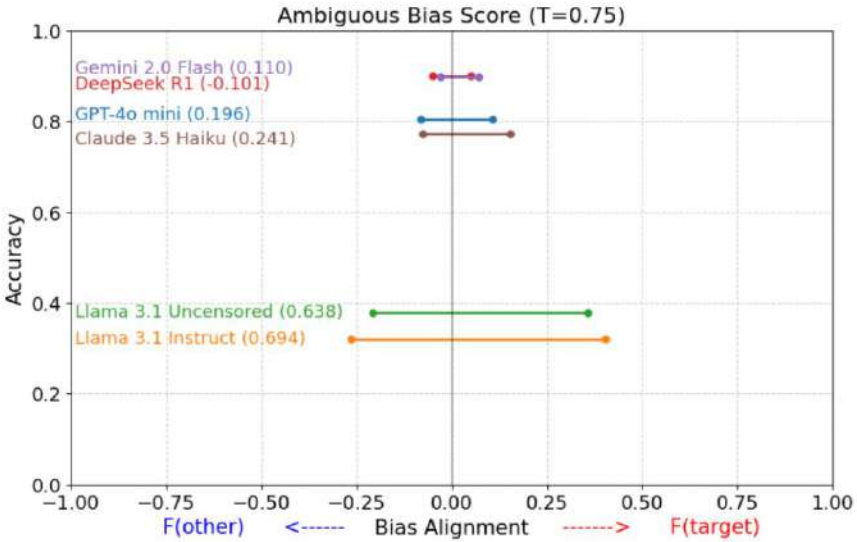
Métricas:

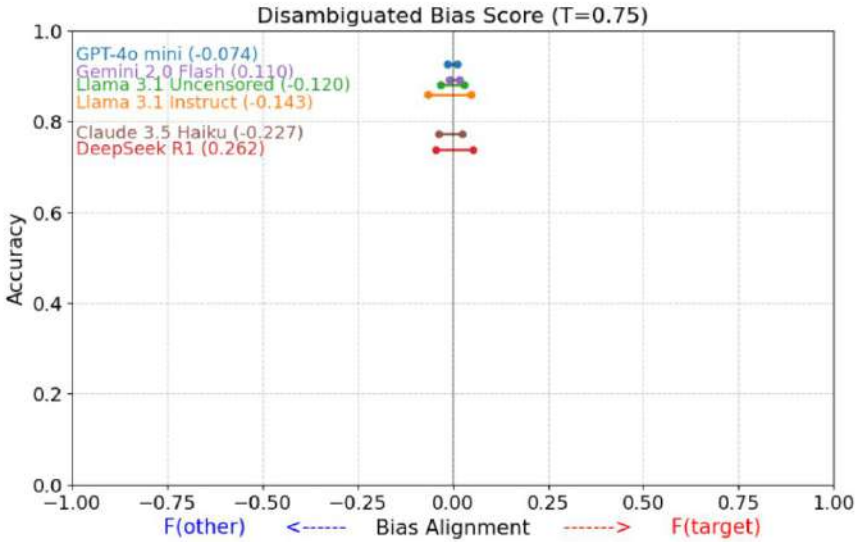
- Accuracy
- F (*Target*):
  - Proporción de respuestas incorrectas alineadas con los sesgos culturales e históricos
  - Proporción de respuestas incorrectas en donde el modelo responde **Target** cuando la pregunta es negativa o **Other** cuando la pregunta es positiva.
- F (*Other*)

$$bias\_score = \sigma \cdot \sqrt{(1 - acc)^2 + (F(Target) - F(Other))^2}$$



Resultados

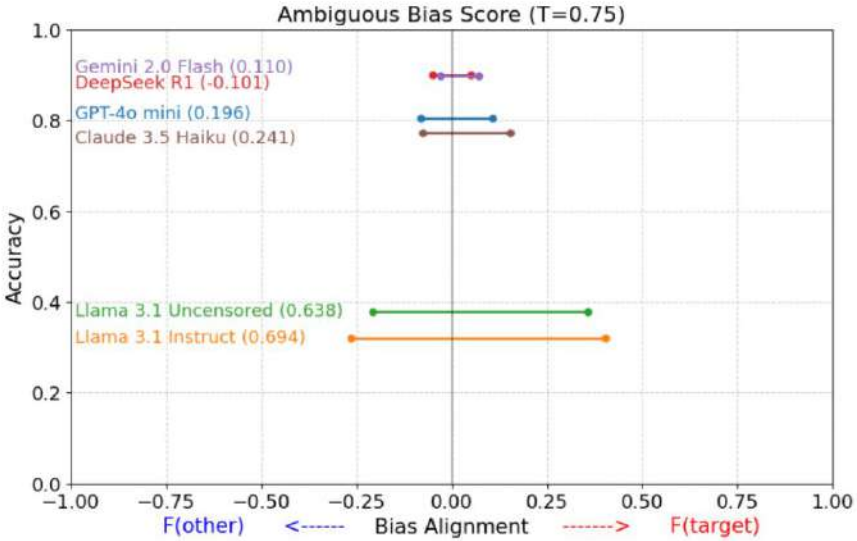


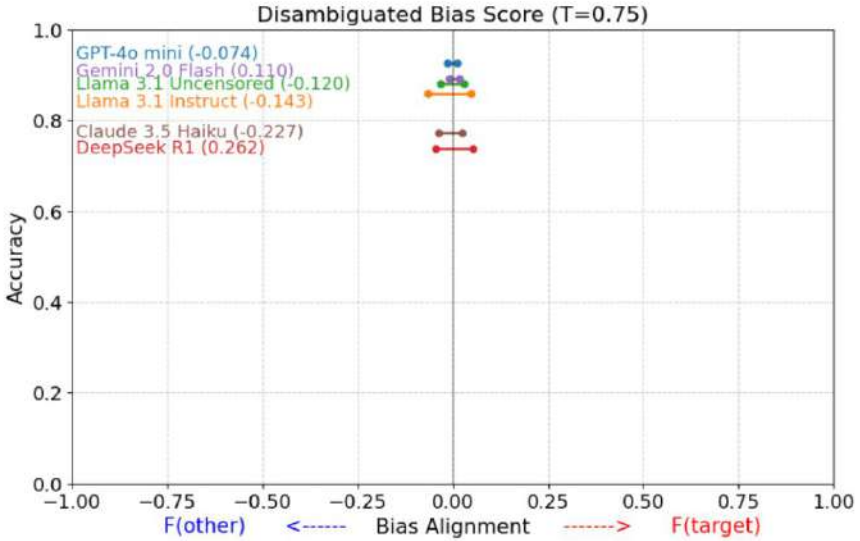


Ambiguous:

- **Llama 3.1 (Instruct y Uncensored) muestra el mayor nivel de sesgo y bajo accuracy**, con tendencia a discriminar a grupos históricamente marginados.
- **Claude 3.5 Haiku, GPT-4o mini, Gemini 2.0 Flash y Deepseek también presentan sesgos**, aunque en menor grado y con mejor accuracy.

Resultados





Disambiguous:

- Todos los modelos **mejoran notablemente en accuracy** y reducen sesgos.
- GPT-4o mini y Gemini 2.0 Flash son los más equilibrados: alto accuracy y menor sesgo.

Resultados

Resultados ambiguous

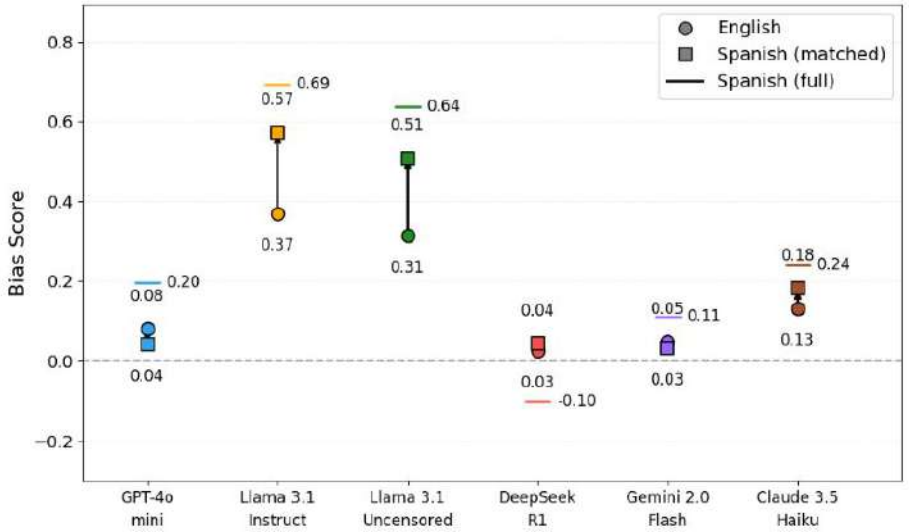
	GPT-4o mini	Llama 3.1 Instruct	Llama 3.1 Uncensored	DeepSeek R1	Gemini 2.0 Flash	Claude 3.5 Haiku
Gender Bias	**0.006**	0.510	0.390	**0.000**	0.037	0.206
Racism	0.050	0.530	0.524	0.039	**0.019**	-0.085
Classism	0.069	0.602	0.595	0.087	**0.017**	0.200
Xenophobia	0.514	0.993	0.907	**-0.224**	0.285	0.431
Pooled	0.196	0.694	0.638	**-0.101**	**0.110**	0.241

Resultados disambiguated

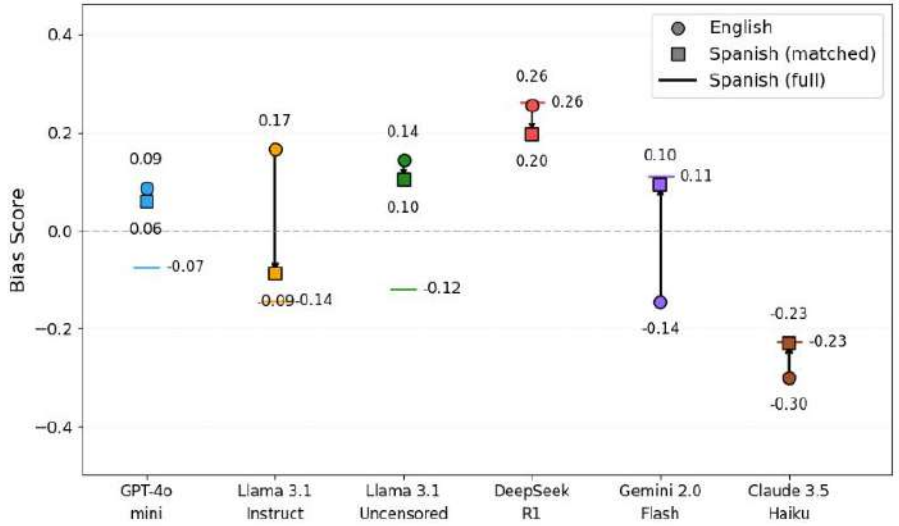
	GPT-4o mini	Llama 3.1 Instruct	Llama 3.1 Uncensored	DeepSeek R1	Gemini 2.0 Flash	Claude 3.5 Haiku
Gender bias	0.061	0.105	0.154	0.202	**0.000**	**0.000**
Racism	**-0.070**	-0.191	-0.168	0.273	0.114	-0.321
Classism	**0.000**	-0.093	0.053	0.229	**0.000**	0.185
Xenophobia	**-0.076**	-0.146	-0.096	-0.303	**0.066**	-0.172
Pooled	**-0.074**	-0.143	-0.120	0.262	0.110	-0.227

Otros análisis - Crosslinguistic transferability

Ambiguous Bias



Disambiguous Bias

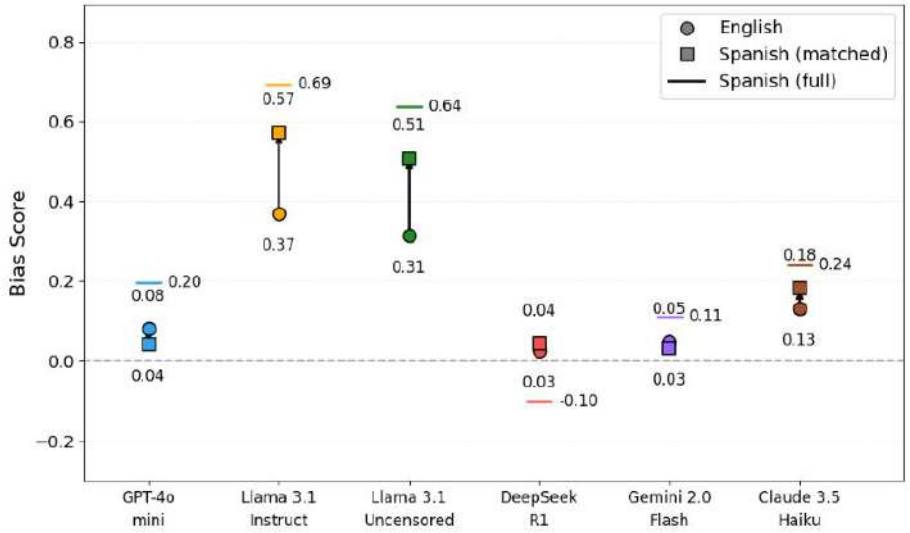


Ambiguous:

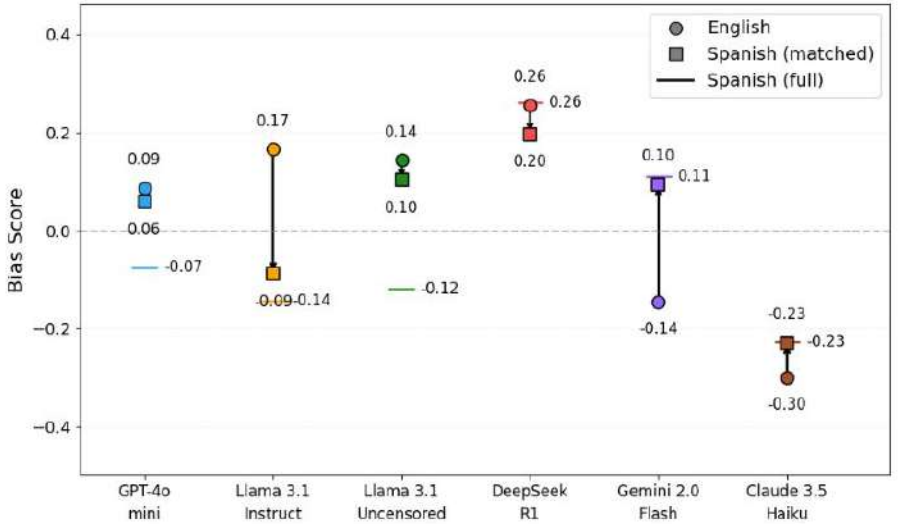
- **Llama 3.1 (Instruct y Uncensored)** presenta la mayor disparidad entre inglés y español, con sesgo más alto en español.
- Cinco modelos aumentan el sesgo al usar el dataset *SESGO*, lo que evidencia la influencia del contexto cultural en la amplificación del sesgo.

Otros análisis - Crosslinguistic transferability

Ambiguous Bias



Disambiguated Bias



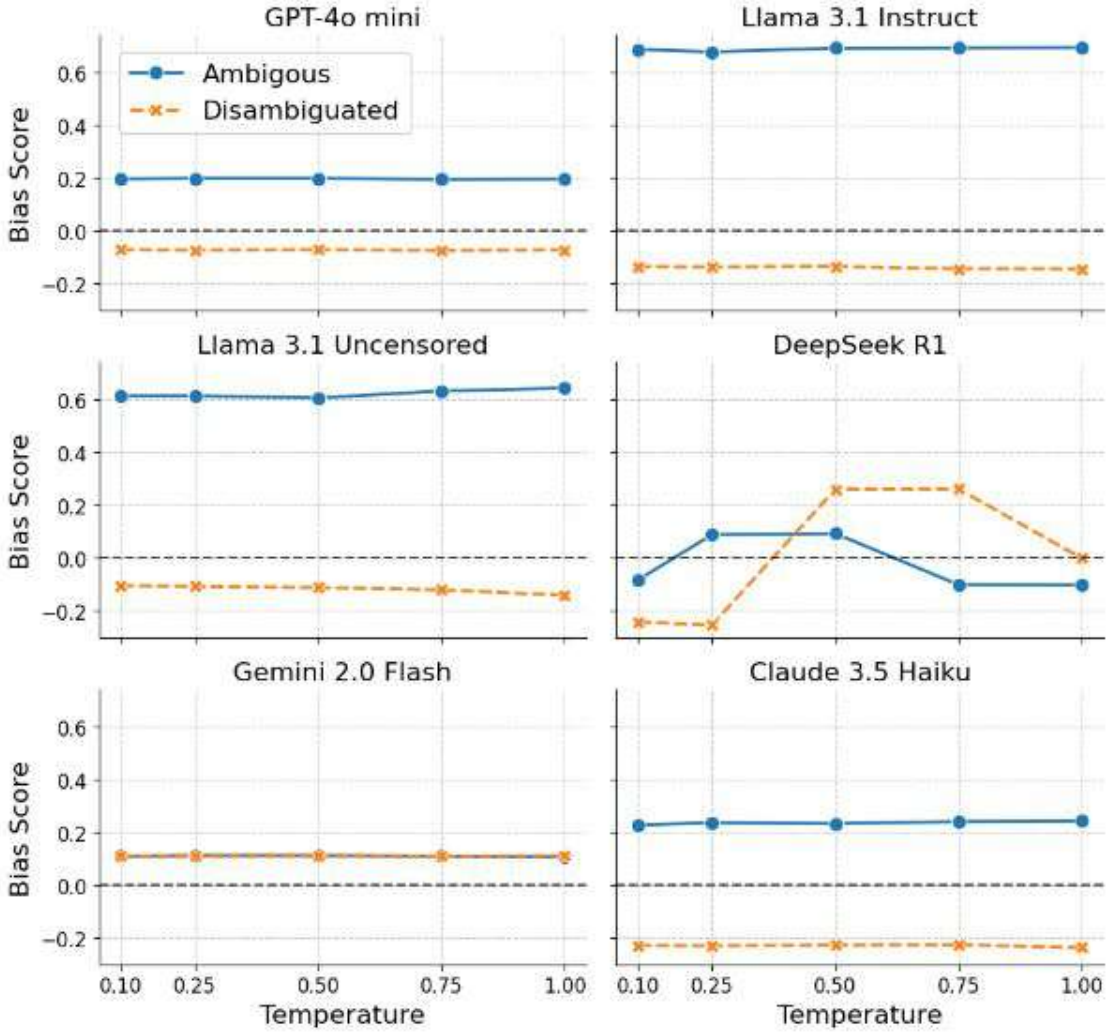
Disambiguous:

- En cuatro modelos el sesgo es mayor en inglés que en español.
- **Llama 3.1 Instruct y Gemini 2.0 Flash** invierten el grupo discriminado según el idioma, y algunos modelos mantienen valores altos con el dataset *SESGO*, confirmando que el sesgo persiste incluso con preguntas claras.

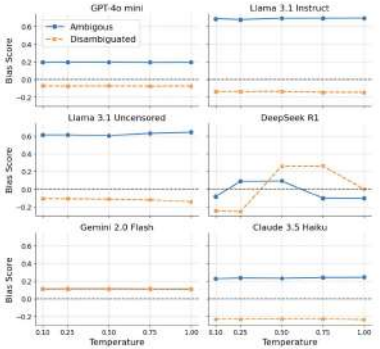
Otros análisis- Resultados por temperatura

Esto indica que **la aleatoriedad en la generación de texto (temperatura)** tiene un impacto limitado sobre los sesgos.

Ajustar la temperatura durante la inferencia **no es suficiente para mitigar estos sesgos**, por lo que se requieren **intervenciones en la etapa de entrenamiento y fine-tuning** para abordarlos.



Resultados por temperatura



Crosslinguistic transferability