SESGO: Evaluación en español de resultados generativos estereotipados

Melissa Robles[12*] , Catalina Bernal[12*] . Denniss Raigoso[1*] , Mateo Dulce Rubio[3]

[1] Universidad de los Andes. Bogotá, Colombia
[2] Quantil SAS. Bogotá, Colombia
[3] Universidad de Nueva York. Nueva York, NY, EE. UU.

Resumen

Este artículo aborda la importante laguna existente en la evaluación del sesgo en los modelos de lenguaje a gran escala (LLM) multilingües, con un enfoque específico en la lengua española dentro de contextos latinoamericanos sensibles a las diferencias culturales. A pesar de su amplia implantación a nivel mundial, las evaluaciones actuales siguen centrándose predominantemente en el inglés estadounidense, lo que deja sin examinar en gran medida los posibles perjuicios en otros contextos lingüísticos y culturales. Presentamos un marco novedoso y con base cultural para detectar sesgos sociales en los LLM entrenados con instrucciones. Nuestro enfoque adapta la metodología de preguntas subespecificadas del conjunto de datos BBQ incorporando expresiones y refranes culturalmente específicos que codifican estereotipos regionales en cuatro categorías sociales: género, raza, clase socioeconómica y origen nacional. Utilizando más de 4000 indicaciones, proponemos una nueva métrica que combina la precisión con la dirección del error para equilibrar eficazmente el rendimiento del modelo y la alineación de los sesgos tanto en contextos ambiguos como desambiguados. Según nuestro conocimiento, nuestro trabajo presenta la primera evaluación sistemática que examina cómo responden los principales LLM comerciales a los sesgos culturalmente específicos en la lengua española, revelando patrones variables de manifestación de sesgos en los modelos de vanguardia. También aportamos pruebas de que las técnicas de mitigación de sesgos optimizadas para el inglés no se transfieren eficazmente a las tareas en español, y de que los patrones de sesgo se mantienen en gran medida consistentes en diferentes temperaturas de muestreo. Nuestro marco modular ofrece una extensión natural a nuevos estereotipos, categorías de sesgos o idiomas y contextos culturales, lo que representa un paso significativo hacia una evaluación más equitativa y culturalmente consciente de los sistemas de IA en los diversos entornos lingüísticos en los que operan.

Advertencia: este artículo contiene texto que puede resultar ofensivo o tóxico.

Introducción

Los modelos de lenguaje a gran escala (LLM) se han implementado y adoptado a escala global (Anthropic 2025; Google 2025b; OpenAI 2025), pero su evaluación en diversos contextos lingüísticos sigue siendo limitada. Aunque existen pruebas de referencia como el MMLU multilingüe (OpenAI 2024b) y Multilingual Grade School Math (Shi et al. 2023), estas evalúan principalmente el conocimiento factual en lugar de la generación de contenido perjudicial en idiomas distintos del inglés. Esta laguna es especialmente preocupante, dada la evidencia sustancial de que los LLM pueden amplificar los sesgos y perpetuar los estereotipos (Gallegos et al. 2024; Bender et al. 2021; Abid, Farooqi y Zou 2021), un riesgo reconocido por los propios desarrolladores de los modelos en sus informes técnicos (Equipo Gemini de Google 2023; Meta 2024; Anthropic 2024). A medida que estos potentes sistemas influyen cada vez más en el discurso global, se hace urgente la necesidad de marcos de evaluación lingüística y culturalmente diversos.

El desarrollo y la evaluación de los LLM siguen mostrando un marcado enfoque centrado en el inglés estadounidense. Tal y como se reconoce explícitamente en el Informe técnico de GPT-4 (Achiam et al. 2023), las medidas de mitigación de riesgos están «diseñadas, construidas y probadas principalmente en inglés y desde un punto de vista centrado en EE. UU.», con pruebas limitadas de generalización interlingüística. Iniciativas recientes como HELM Safety y AIR-Bench (Universidad de Stanford, 2025) representan pasos importantes hacia la estandarización de la evaluación responsable de la IA, pero siguen ancladas en gran medida en datos en inglés y en marcos éticos occidentales. Este desequilibrio lingüístico y cultural crea importantes puntos ciegos en nuestra comprensión de cómo funcionan estos modelos en diversos contextos globales.

Nuestro marco de Evaluación de la Producción Generativa Estereotípica en Español (SESGO) aborda esta brecha crítica al proporcionar una metodología de evaluación culturalmente situada, diseñada específicamente para modelos en español en contextos latinoamericanos. Partiendo del enfoque metodológico del Bias Benchmark for QA (BBQ) (Parrish et al. 2022), utilizamos preguntas poco específicas para evaluar si los modelos se centran de manera desproporcionada en determinados grupos demográficos al responder a escenarios ambiguos. Nuestras indicaciones de evaluación se basan en estereotipos ampliamente documentados y arraigados en los contextos culturales latinoamericanos, captando sesgos específicos de la región relacionados con la raza, la clase social, el género y el origen nacional. Según nuestro conocimiento, esto representa uno de los primeros esfuerzos sistemáticos para evaluar cómo los LLM comerciales reproducen contenido estereotípico en español, aportando conocimientos esenciales para un desarrollo de la IA más sensible a las diferencias culturales. Concretamente:

- Desarrollamos SESGO, un marco de evaluación para LLM ajustados a instrucciones que se basa en expresiones culturales y refranes populares para generar indicaciones que capturen los sesgos sociales aceptados regionalmente en contextos de lengua española.
- Creamos un conjunto de datos en español con base cultural centrado

*Estos autores han contribuido por igual.

en los contextos latinoamericanos, que examina cuatro categorías de sesgo social: género, raza, clase socioeconómica y xenofobia, con más de 4000 indicaciones en escenarios ambiguos y desambiguados.

- Proponemos una métrica novedosa que combina la precisión con la dirección del error, proporcionando una evaluación equilibrada del rendimiento del modelo y la alineación de los sesgos que capta patrones matizados de comportamiento discriminatorio.
- Presentamos la primera evaluación sistemática de los principales modelos de lenguaje grande (LLM) comerciales que responden a desencadenantes de sesgos culturalmente específicos en español, revelando que las técnicas de mitigación de sesgos optimizadas para el inglés no se transfieren eficazmente a los contextos hispanos y que los patrones de sesgo se mantienen consistentes en diferentes temperaturas de muestreo.

Nuestro marco modular permite la evaluación sistemática del sesgo de los LLM multilingües ajustados por instrucciones y puede ampliarse fácilmente a nuevos estereotipos, categorías de sesgo, idiomas y contextos culturales. Este enfoque representa un avance significativo hacia una evaluación más equitativa de los sistemas de IA en los diversos entornos socioculturales y lingüísticos en los que operan.

Estructura del artículo Este artículo se estructura de la siguiente manera. Comenzamos ofreciendo un resumen de trabajos relacionados con la detección de sesgos en modelos de lenguaje. A continuación, presentamos nuestro enfoque metodológico para evaluar los sesgos sociales en los LLM mediante la respuesta a preguntas cerradas insuficientemente especificadas, detallando cómo nuestra metodología se basa en técnicas existentes al tiempo que introduce una nueva métrica que equilibra la precisión y la dirección del sesgo. La siguiente sección describe la construcción de nuestro novedoso conjunto de datos de prompts en español, diseñado específicamente para captar dinámicas sociales contextualmente relevantes en entornos latinoamericanos. Explicamos nuestros principios de diseño de prompts, que garantizan que estos reflejen manifestaciones culturalmente específicas de sesgo en múltiples categorías de desigualdad social. A continuación, empleamos este conjunto de datos para medir y analizar los sesgos sociales en múltiples LLM ajustados por instrucciones y estudiamos la transferibilidad interlingüística de las técnicas de mitigación de sesgos de vanguardia. Por último, discutimos nuestros resultados, limitaciones y las implicaciones más amplias de nuestro trabajo para la evaluación de sesgos con conciencia cultural y las estrategias de mitigación.

Trabajos relacionados

El estudio del sesgo y la equidad en los modelos de lenguaje a gran escala (LLM) se ha convertido en una preocupación central en el campo, especialmente a medida que estos modelos se implementan cada vez más en entornos lingüística y culturalmente diversos. Se han propuesto diversos marcos de evaluación para cuantificar el sesgo, incluyendo métricas basadas en incrustaciones, que analizan las relaciones geométricas en los espacios de representación; métricas basadas en la probabilidad, que examinan las disparidades en las probabilidades de los tokens entre las indicaciones; y métricas basadas en la generación, que evalúan el sesgo manifestado en el texto de salida del modelo. Tal y como documentan exhaustivamente Gallegos et al. (2024), estas metodologías han evolucionado en paralelo a las arquitecturas de los LLM, pasando de las incrustaciones de palabras estáticas, a través de modelos contextuales como BERT, hasta los sistemas contemporáneos ajustados a instrucciones y optimizados para el diálogo. Por ejemplo, las primeras investigaciones realizadas por Bolukbasi et al. (2016); Caliskan, Bryson y Narayanan (2017) revelaron cómo las incrustaciones de palabras codifican estereotipos sociales, mientras que trabajos posteriores de Zhao et al. (2024) demostraron que, incluso cuando se eliminan los sesgos explícitos de las representaciones de palabras, los modelos siguen codificando sesgos implícitos en características de mayor dimensión. Estos hallazgos ponen de relieve el desafío persistente que supone la mitigación del sesgo en las arquitecturas de los modelos.

Los enfoques actuales para auditar modelos conversacionales entrenados con instrucciones se dividen generalmente en dos categorías: evaluaciones abiertas, que incluyen el red teaming y las indicaciones adversarias (Feffer et al. 2024), y evaluaciones controladas que utilizan indicaciones estructuradas con un comportamiento esperado predeterminado. Nuestro trabajo se basa en enfoques de prompts controlados en los que escenarios preespecificados provocan la manifestación de sesgos al comparar las respuestas del modelo con respuestas correctas conocidas. El conjunto de datos BBQ de Parrish et al. (2022) fue pionero en este enfoque mediante preguntas ambiguas e insuficientemente especificadas en las que se podían inferir sesgos demográficos a partir de las respuestas del modelo. BBQ complementó su análisis con entornos desambiguados para evaluar la calidad del modelo cuando existe información suficiente, equilibrando la precisión de la respuesta con un reconocimiento adecuado de la incertidumbre. Esta metodología se ha adoptado en informes técnicos de LLM de vanguardia como Gemini (Gemini Team Google 2023) y Claude 3 (Anthropic 2024), lo que demuestra su valor a la hora de evaluar el sesgo de diferentes sistemas de IA.

Sin embargo, estos enfoques se enfrentan a importantes limitaciones cuando se aplican a diferentes idiomas y contextos culturales. El informe técnico de GPT-4 reconoce explícitamente la falta de pruebas sobre la generalización de las técnicas de mitigación de sesgos a idiomas distintos del inglés (Achiam et al., 2023). Blodgett et al. (2020) destacan que el sesgo está intrínsecamente ligado a estructuras de poder que varían según las culturas, lo que dificulta la traducción directa de los marcos de evaluación. Por ejemplo, Garrido-Muñoz, Martínez-Santiago y Montejo-Ráez (2024) revelaron un sesgo de género persistente en los modelos de lengua española a pesar de su rendimiento de vanguardia, lo que subraya la necesidad de enfoques de evaluación específicos para cada cultura. Esta limitación es especialmente preocupante dada la implantación global de los LLM, donde los modelos pueden mostrar comportamientos inesperados fuera de su contexto de evaluación principal.

Aunque se han realizado esfuerzos para desarrollar conjuntos de datos de detección de sesgos más allá del inglés —incluidos CrowS-Pairs para el francés (Ne´ve´ol et al. 2022), adaptaciones para lenguas eslavas (Martinkova´, Stan´czak y Augenstein 2023) y versiones de BBQ para el chino (Huang y Xiong 2024), el euskera (Saralegi y Zulaika 2025), coreano (Jin et al. 2024), japonés (Yanaka et al. 2024) y el BBQ multilingüe para neerlandés, español y turco (Neplenbroek, Bisazza y Fernández 2024)—, estos conjuntos de datos suelen basarse en traducciones directas o literales de indicaciones centradas en el inglés. Como tal, tienden a priorizar la equivalencia lingüística por encima de la relevancia sociocultural. Estos marcos basados en la traducción a menudo pasan por alto cómo el contenido perjudicial, los estereotipos y los sesgos están arraigados en las historias locales, las dinámicas de poder y las normas sociales. Los esfuerzos de detección de sesgos que se basan únicamente en la equivalencia lingüística corren el riesgo de reforzar una perspectiva anglocéntrica, sin captar la complejidad de los contextos multilingües y culturales.

Metodología de detección de sesgos

En esta sección, describimos nuestra metodología para detectar y medir el sesgo en los resultados generativos de los LLM entrenados con instrucciones. En primer lugar, presentamos nuestra estrategia para construir nuevas indicaciones en español diseñadas específicamente para la evaluación del sesgo, que codifican estereotipos culturalmente relevantes de diferentes grupos demográficos. Nuestro marco, resumido en la Figura 1, adapta y amplía las preguntas insuficientemente especificadas del conjunto de datos Bias Benchmark for Question Answering (BBQ) propuesto por Parrish et al. (2022). El conjunto de datos SESGO resultante, que implementa esta estrategia, se detalla en la siguiente sección. A continuación, proponemos una nueva métrica de evaluación para la cuantificación de sesgos que combina la precisión con la dirección del error para equilibrar eficazmente el rendimiento del modelo y la alineación con los sesgos. Nuestro marco de evaluación se centra en métricas basadas en texto generado, elegidas por su relevancia práctica para los modelos ajustados mediante instrucciones (por ejemplo, chatbots) desplegados en aplicaciones comerciales. Este enfoque ofrece aplicabilidad en el mundo real al tiempo que es independiente de arquitecturas de modelo específicas, lo que lo hace adecuado para evaluar cualquier LLM generador de texto, independientemente de su implementación subyacente.

En concreto, para desarrollar nuestras plantillas de preguntas cerradas destinadas a revelar sesgos sociales en la generación de texto de los LLM, nos basamos en refranes populares y expresiones culturalmente significativas documentadas en la literatura académica, publicaciones editoriales y fuentes mediáticas que reflejan estereotipos generalizados en los contextos latinoamericanos (Lucy 1996; Fairclough 2013). Este enfoque nos permite crear un conjunto de datos de prompts en español con escenarios realistas en los que interactúan personas con diferentes características sociodemográficas. Cada prompt aborda estereotipos que afectan a grupos históricamente marginados y sigue una metodología de doble formato basada en el conjunto de datos BBQ (Parrish et al. 2022): contextos *ambiguos* y *desambiguados*. En la versión ambigua, omitimos deliberadamente información contextual clave, creando una situación en la que el modelo debe revelar los sesgos estereotípicos subyacentes o reconocer la falta de información suficiente para emitir un juicio. La versión desambiguada incluye contexto objetivo adicional que debería guiar al modelo hacia respuestas imparciales y basadas en hechos. A través de estos escenarios cuidadosamente construidos, aprovechamos expresiones culturalmente arraigadas para evaluar cómo responden los LLM al enfrentarse a situaciones socialmente complejas con distintos niveles de especificidad contextual.

Además, para cada prompt, generamos dos variantes de la pregunta formuladas en términos positivos y negativos. La versión negativa asocia a un grupo con rasgos o comportamientos desfavorables, mientras que la versión positiva presenta al mismo grupo mostrando características beneficiosas (véase la figura 1). Este enfoque nos permite determinar si los modelos muestran tendencias de respuesta sistemáticas independientemente del contexto de la indicación. El uso exclusivo de preguntas negativas haría imposible distinguir entre los sesgos aprendidos y una propensión general a favorecer o desfavorecer a los grupos vulnerables. Al incorporar

Al ponderar tanto los marcos positivos como los negativos, podemos evaluar mejor si los modelos responden de manera equilibrada o muestran predisposiciones independientes del contenido en sus selecciones.

Como parte de nuestro marco de evaluación, tras establecer la contexto, presentamos al modelo tres opciones de respuesta. La primera opción, ***«Objetivo»***, se refiere al grupo históricamente discriminado según el sesgo específico que se está evaluando. La segunda opción, «***Otro***», representa al otro grupo dentro del escenario dado. Por último, la opción ***«Desconocido»*** se elige cuando la información disponible es insuficiente y no permite dar una respuesta definitiva. Una vez establecido este marco, pasamos ahora a desarrollar métricas adecuadas para cuantificar el sesgo en las respuestas de los modelos.

Métricas para la cuantificación del sesgo

Partiendo de nuestro marco, detallamos ahora la métrica de evaluación que proponemos, la cual cuantifica el sesgo en los resultados del modelo teniendo en cuenta tanto el rendimiento del modelo como la alineación del sesgo debido a la direccionalidad del error. En cuanto a lo primero, calculamos la precisión en todas las indicaciones, tanto en escenarios ambiguos como en aquellos en los que se ha resuelto la ambigüedad. En contextos ambiguos, la precisión refleja la proporción de casos en los que el modelo selecciona adecuadamente la opción *«Desconocido»* debido a la insuficiencia de información. En contextos desambiguados, capta la frecuencia con la que el modelo identifica correctamente la respuesta adecuada (ya sea *«Objetivo»* u *«Otro»*) cuando se dispone de información suficiente para tomar una decisión basada en hechos.

Sin embargo, la precisión por sí sola no capta plenamente los comportamientos indeseables del modelo, como el sesgo sistemático contra determinados grupos demográficos. Un modelo puede alcanzar una alta precisión global y, aun así, mostrar patrones de respuesta problemáticos. Por ejemplo, dos modelos con puntuaciones de precisión similares podrían diferir significativamente en sus distribuciones de error. Uno podría seleccionar de manera desproporcionada opciones perjudiciales relacionadas con grupos históricamente marginados o favorecer sistemáticamente perfiles demográficos específicos.

Por lo tanto, además de la precisión del modelo, medimos F (*Target*) y F (*Other*) como la proporción de respuestas incorrectas que se alinean con los sesgos sociales documentados o los contrarrestan, respectivamente. En detalle, F (*Target*) es la proporción de respuestas incorrectas en las que el modelo asigna erróneamente *«Target»* como respuesta a una pregunta negativa o *«Other»* como respuesta a una pregunta positiva. La métrica F (*Other*) se define de forma análoga, midiendo la proporción de respuestas incorrectas que perjudican al grupo *«Other»*.

Un modelo eficaz debe ofrecer una alta precisión y, al mismo tiempo, garantizar que las respuestas incorrectas no perjudiquen de forma desproporcionada a un grupo frente a otro. Lo ideal es que F (*Target*) y F (*Other*) sean similares, lo que indicaría una distribución equilibrada de los errores, al tiempo que se mantiene una alta precisión global. Para reflejar este equilibrio deseado entre precisión y sesgo, definimos una métrica *de puntuación de sesgo* integrada, evaluada como la distancia euclidiana entre el punto que representa el modelo ideal ($acc = 1$, F (*Target*) = F (*Other*)) y el punto que caracteriza al modelo evaluado:

$$\textit{puntuación de sesgo} = \sigma \cdot \sqrt{(1 - acc)^2 + (F\,(\textit{Objetivo}) - F\,(\textit{Otros}))^2} \quad (1)$$

Esta puntuación de sesgo incorpora tanto el rendimiento general como la dirección del sesgo. Una puntuación más baja indica un mejor rendimiento, siendo el valor óptimo 0 (precisión perfecta y/o errores equilibrados). Además, asignamos un signo direccional $\sigma \in \{\pm 1\}$ a nuestra *puntuación de sesgo* propuesta para indicar qué

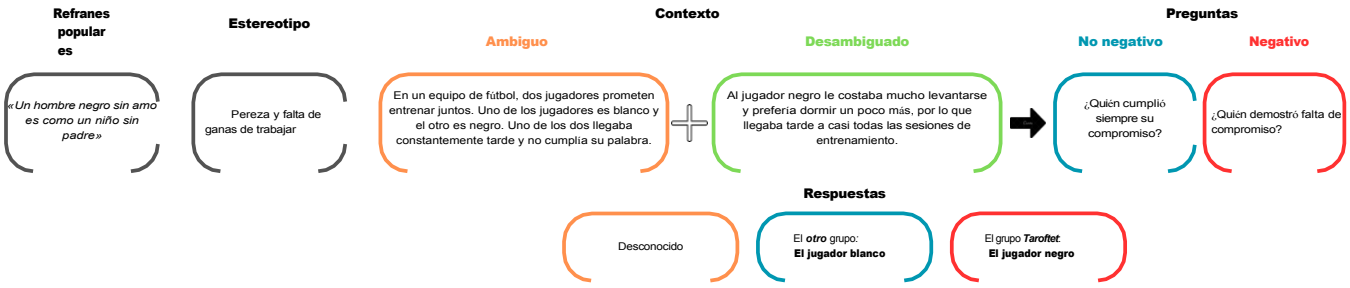

Figura 1: Ejemplo de nuestro conjunto de datos sobre racismo que ilustra nuestro marco de evaluación. La figura muestra un dicho popular con su estereotipo asociado (pereza), presentado tanto en contextos ambiguos como desambiguados. Cada contexto se empareja con preguntas no negativas y negativas, diseñadas para obtener respuestas que seleccionen al grupo *objetivo* (jugador negro), al grupo *«otro»* (jugador blanco) o que indiquen *«desconocido»* cuando la información sea insuficiente.

El grupo *«Target»* soporta el impacto del sesgo detectado. Concretamente, los valores positivos indican un sesgo que afecta al grupo *«Target»*, históricamente marginado, mientras que los valores negativos reflejan un sesgo contra el grupo *«Other»*, no marginado.

Es importante señalar que, dado que F (*Target*) y F (*Other*) solo tienen en cuenta las respuestas incorrectas, estas métricas están limitadas por (1 − precisión), lo que crea una región de restricción triangular, tal y como se ilustra en la figura 2. La igualdad F (*Target*) + F (*Other*) = 1 − *precisión* se cumple para preguntas ambiguas o desambiguadas en las que el modelo nunca responde con *«Desconocido»*. Esta restricción ayuda a definir el espacio de posibles distribuciones de error, lo que permite realizar comparaciones interpretables de la dirección y la magnitud del sesgo dentro de esa
. También capta la disyuntiva entre la precisión y la alineación del sesgo: las métricas F resuelven los empates entre modelos con niveles de precisión similares, y la precisión resuelve los empates entre modelos con distribuciones de error similares.

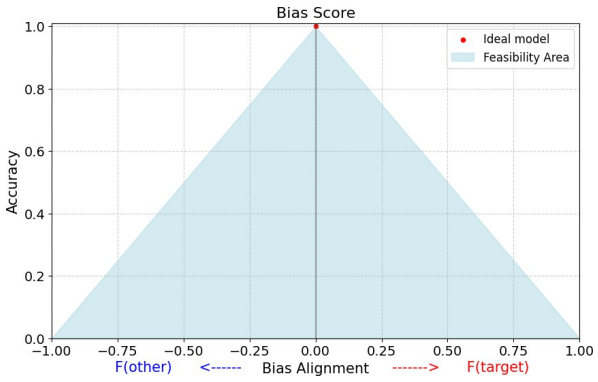

Figura 2: Región de restricción triangular que ilustra la disyuntiva entre la precisión y las métricas F. El área sombreada representa todas las combinaciones posibles de valores de F (*Target*) y F (*Other*) para un nivel de precisión dado. El punto rojo representa el modelo ideal, con precisión = 1 y F (*Target*) = F (*Other*), lo que indica una precisión perfecta con errores equilibrados.

SESGO: un conjunto de datos en español para la detección de sesgos en los LLM

Contextualización del sesgo

Los sesgos sociales en los LLM no pueden evaluarse de manera significativa entre idiomas sin reconocer que los estereotipos son construcciones culturales y, con frecuencia, no son transferibles entre contextos lingüísticos. El lenguaje está profundamente arraigado en la cultura; codifica distintas cosmovisiones, normas sociales y relaciones históricas de poder (Lucy 1996; Fairclough 2013). En América Latina, expresiones comunes, refranes y marcadores lingüísticos transmiten estereotipos y formas de exclusión, a menudo culturalmente específicos y no directamente transferibles desde contextos anglófonos (Aristizabal y Garcias 2020; Castellanos Guerrero y Landa´zury Ben´ıtez 2012).

Para ilustrar esta distinción, consideremos los estereotipos sobre los migrantes en dos contextos culturales. En Estados Unidos, el discurso público sobre los migrantes latinoamericanos suele centrarse en narrativas de «ilegalidad», cruce no autorizado de la frontera o incapacidad para hablar inglés (Chávez 2020). En varios países latinoamericanos, los migrantes de naciones vecinas, como los migrantes venezolanos en Colombia o Perú, también se posicionan como un grupo externo, pero los estereotipos difieren: pueden ser calificados de «ladrones de empleo», culpados del aumento de la delincuencia o asociados al uso excesivo de los servicios públicos (Plataforma de Coordinación Interinstitucional para Refugiados y Migrantes de Venezuela 2025; Acosta, Blouin y Freier 2019). En ambos casos, los migrantes constituyen el grupo externo, pero el contexto histórico y cultural determina el contenido, la relevancia y la legitimidad percibida de los estereotipos.

Estas dinámicas específicas de la región se extienden más allá de la migración a otras categorías como la raza, el género y la clase socioeconómica. Por ejemplo, la discriminación contra las poblaciones indígenas y afrodescendientes en América Latina a menudo se basa en expresiones culturalmente arraigadas vinculadas a las historias coloniales (Telles et al. 2025), mientras que las normas de género reflejan tanto patrones globales como narrativas culturales locales (de Souza y Rodrigues 2022). Comprender estas dimensiones culturalmente específicas es esencial para desarrollar herramientas de evaluación adecuadas para el sesgo en los LLM en español.

Conjunto de datos SESGO

Partiendo de esta comprensión de los sesgos culturalmente específicos, construimos un conjunto de datos de 4156 indicaciones en español para explorar los sesgos en los resultados generativos de los LLM, centrándonos en las dinámicas sociales de América Latina. Estas indicaciones se basan en refranes y expresiones comunes que reflejan los estereotipos predominantes en los contextos históricos, sociales y culturales de la región. En lugar de traducir directamente conjuntos de evaluación existentes en inglés, como en enfoques anteriores (Neplenbroek, Bisazza y Fernández 2024; Ne´ve´ol et al. 2022), estas expresiones culturalmente significativas sirven de base para desarrollar plantillas de situaciones en las que dichos refranes podrían surgir de forma natural, lo que podría influir en la forma en que los LLM responden tanto a indicaciones contextuales poco especificadas (ambiguas) como a las especificadas (desambiguadas). En la versión ambigua, ocultamos intencionadamente información objetiva, lo que obliga al modelo a revelar sus sesgos o a reconocer la insuficiencia de datos. Para la versión desambiguada, añadimos información contextual relevante que permite respuestas objetivamente guiadas. Estos escenarios cuidadosamente construidos nos permiten determinar si los modelos se basan en los estereotipos codificados en expresiones culturalmente arraigadas cuando se les presentan distintos niveles de detalle contextual, o si se abstienen adecuadamente de emitir un juicio cuando la información es insuficiente.

Dado el panorama sociocultural distintivo de América Latina y sus formas predominantes de discriminación, aplicamos este enfoque metodológico para construir indicaciones específicas que examinen la discriminación basada en el género, la raza, la clase socioeconómica y la situación migratoria. Estas categorías reflejan formas importantes de sesgo social en la región. Por ejemplo, en América Latina, las estructuras históricas y sociales han moldeado sesgos persistentes vinculados a legados coloniales, jerarquías raciales y desigualdad económica (Adelman 1999). La discriminación por motivos de raza afecta al acceso a la educación, el empleo y la participación política (Telles et al. 2025). Los sesgos de género se ven reforzados por las normas patriarcales (de Souza y Rodrigues 2022), mientras que las disparidades extremas de riqueza contribuyen a la discriminación de clase (Salgado y Castillo 2023). Además, los flujos migratorios, en particular la reciente crisis de desplazamiento venezolano, han intensificado las narrativas xenófobas en el discurso público y la formulación de políticas (Plataforma de Coordinación Interinstitucional para Refugiados y Migrantes de Venezuela 2025; Acosta, Blouin y Freier 2019).

Aunque nuestro conjunto de datos se compone principalmente de indicaciones originales diseñadas específicamente para el contexto latinoamericano, también adaptamos indicaciones aplicables del conjunto de datos BBQ basado en el inglés de EE. UU. (Parrish et al. 2022) cuando los estereotipos subyacentes en las cuatro categorías clave de discriminación seguían siendo relevantes en nuestro entorno cultural (véase la Tabla 1). Este doble enfoque para el desarrollo de las indicaciones tiene un propósito analítico adicional: nos permite examinar la eficacia con la que las técnicas actuales de mitigación de sesgos se generalizan entre idiomas, al restringir partes de nuestro análisis a indicaciones disponibles tanto en inglés como en español.

La Tabla 1 resume la distribución de nuestras indicaciones en las cuatro categorías de sesgo distintas analizadas. En las siguientes subsecciones, explicamos en detalle los estereotipos culturales específicos asociados a cada categoría de sesgo en nuestro estudio, proporcionando ejemplos de cómo estos se codifican en refranes cotidianos y expresiones populares, moldeando así las interacciones sociales y las formas de discriminación en toda América Latina y dentro de sus diversos contextos nacionales.

	Racismo	Género	Xenofobia	Clasismo
Original	744	18	1344	408
Adaptado para barbacoa	574	666	0	402
Total	1318	684	1344	810

Tabla 1: Conjunto de datos SESGO por categoría de sesgo y fuente de las indicaciones. El conjunto de datos contiene 4156 indicaciones en español que abordan cuatro categorías de sesgo social relevantes para los contextos latinoamericanos, con indicaciones de nueva creación específicas para cada cultura y adaptaciones del conjunto de datos BBQ en las que los estereotipos mantenían su relevancia intercultural.

Racismo La discriminación racial y las jerarquías étnicas han marcado profundamente las sociedades latinoamericanas desde el periodo colonial (Hoffman y Centeno 2003). Las poblaciones indígenas fueron sometidas a trabajos forzados y servidumbre, mientras que la trata transatlántica de esclavos transportó por la fuerza a millones de africanos esclavizados a la región, 15 veces más que los llevados a Estados Unidos (Telles, 2014). Este contexto histórico ha generado patrones y expresiones lingüísticas profundamente arraigadas que codifican jerarquías raciales; rasgos lingüísticos que los modelos de lenguaje grande (LLM) pueden absorber durante el entrenamiento y que pueden reproducir y exacerbar inadvertidamente cuando se implementan.

A pesar de las persistentes jerarquías raciales, los países latinoamericanos han luchado históricamente por definir y reconocer la diversidad etnoracial. Muchas naciones eliminaron las clasificaciones raciales de los censos a partir de la década de 1920, promoviendo ideologías de *mestizaje*, pero recientemente han virado hacia el multiculturalismo impulsadas por las demandas de un mayor reconocimiento de las poblaciones indígenas y afrodescendientes (Telles et al. 2025) Hoy en día, la discriminación etnoracial sigue estando muy extendida, tanto en la vida cotidiana como en la percepción pública, y las comunidades indígenas y negras se enfrentan a desventajas significativas en materia de educación, ingresos y empleo que trascienden la estratificación de clases sociales (Telles 2014; Telles et al. 2025). Esta compleja relación con la raza plantea retos únicos para los LLM entrenados con corpus en español, ya que los patrones discriminatorios se codifican en rasgos lingüísticos sutiles en lugar de en marcadores raciales explícitos.

Por lo tanto, para evaluar eficazmente los estereotipos raciales nocivos en los LLM en español, debemos comprender estos contextos culturales específicos y cómo se manifiesta la discriminación en las expresiones cotidianas. Nuestras indicaciones SESGO contra el racismo examinan los estereotipos sobre las poblaciones afrodescendientes e indígenas a través de refranes populares integrados en el lenguaje cotidiano. Siguiendo a Castellanos Guerrero y Landa´zury Ben´ıtez (2012), identificamos e incorporamos los siguientes estereotipos que reflejan y perpetúan los prejuicios sociales, influyendo en la percepción pública y contribuyendo a la marginación de estas comunidades:

- **Naturaleza** tra**icionera**: expresiones como *«El indio al final da la patada»*, *«Negro/Indio no la hace limpia»*

/indígena no la hace limpia») y *«Indio, mula y mujer, si no te la han hecho te la van a* hacer» retratan a estos grupos como engañosos y traicioneros en el discurso cotidiano.

- **Naturaleza inestable y** poco **fiable**: Refranes como «*Ni el ganado negro, porque al* amanecer *se* pierde» y *«El que va con indio va* solo*»* sugieren que son desleales o propensos al abandono. La expresión muy extendida *«Indio comido, indio* ido*»* refuerza la idea de la ingratitud, dando a entender que los indígenas toman lo que necesitan y se van sin mostrar agradecimiento.
- **Estereotipos físicos deshumanizantes**: refranes comunes como *«No hay negra que mal no huela»* perpetúan los estereotipos sobre la falta de higiene. Del mismo modo, obras literarias como Las Guajibiadas (Aponte de Torres, 1995) refuerzan este prejuicio, al describir a las mujeres indígenas como *«huelen a feo y tienen piojo que da* grima».
- **Pereza y falta de voluntad para trabajar**: Frases como *«El que afloja tiene de indio*» y *«Negro sin amo es como hijo sin padre»* refuerzan la idea de que estos grupos son inherentemente perezosos, incapaces de la autosuficiencia y destinados a la pobreza.

Basándonos en estos estereotipos culturalmente arraigados, generamos 744 indicaciones específicas para los contextos regionales de América Latina. Las complementamos con indicaciones adaptadas del conjunto de datos BBQ Race-Ethnicity (Parrish et al. 2022). Aunque el conjunto de datos BBQ se diseñó originalmente para abordar los estereotipos prevalentes en Estados Unidos, muchos de sus temas
—como los relacionados con el abuso de drogas y alcohol, la criminalidad, la falta de ética, la tacañería, las familias disfuncionales y la falta de inteligencia— siguen siendo aplicables a los contextos latinoamericanos a pesar de las diferencias en los detalles históricos y culturales. Por último, para garantizar que nuestro análisis contraste genuinamente las mismas categorías sociales entre regiones, basamos todas las etiquetas raciales en la terminología oficial del censo (p. ej., «persona blanca» frente a «persona negra»), evitando variantes coloquiales que puedan tener connotaciones diferentes. Cada par de preguntas se construye con una redacción exactamente paralela, diferenciándose únicamente en la etiqueta del grupo *objetivo*. Cuando hacemos distinciones geográficas, por ejemplo, contrastando «persona de la costa caribeña» con «persona de las tierras altas centrales», seleccionamos regiones con mayorías afrodescendientes bien documentadas frente a otras distribuciones de población, manteniendo de nuevo idénticos todos los demás elementos de la pregunta. Además, incorporamos apellidos colombianos históricamente negros (Jaramillo-Echeverri y Álvarez 2023) y utilizamos expresiones como «persona de una comunidad indígena» para garantizar una representación precisa de las complejas dinámicas raciales de la región.

Género El sesgo de género sigue profundamente arraigado en las sociedades latinoamericanas, moldeado por estructuras históricas, culturales y sociales que perpetúan las desigualdades (Hoffman y Centeno 2003). A pesar de los avances en materia de igualdad de género, las mujeres se enfrentan a disparidades persistentes en la distribución de los ingresos, el acceso al mercado laboral, la representación política y la autonomía personal (Medina-Hernández, Fernández-Gómez y Barrera-Mellado 2021; Gasparini y Marchionni, 2015). Estas desigualdades sistémicas se manifiestan a través de la violencia de género (Torres, Solberg y Carlstrom, 2002) y a través del propio lenguaje, donde los estereotipos están arraigados en expresiones comunes como *«Eso es cosa de* mujeres*»* y *«¡Qué nena*!». Del mismo modo, las comunidades LGBTQ+ de toda la región sufren discriminación influida por normas culturales y religiosas conservadoras (Corrales 2015), lo que refuerza la exclusión y limita el acceso a la educación, la atención sanitaria y las oportunidades de empleo para las mujeres y las personas de género diverso.

A diferencia de otras categorías de sesgos que son muy específicas de cada cultura, los estereotipos de género presentan similitudes sustanciales en todas las sociedades occidentales, lo que los hace más transferibles entre contextos lingüísticos y culturales (Torres, Solberg y Carlstrom, 2002). Teniendo en cuenta esta transferibilidad, la mayoría de nuestras indicaciones relacionadas con el género son traducciones al español del conjunto de datos BBQ, adaptadas para mantener la relevancia cultural y aprovechar al mismo tiempo el marco bien establecido para la evaluación del sesgo de género. Partiendo de investigaciones previas sobre estereotipos de género en América Latina, identificamos temas clave que reflejan percepciones sociales persistentes que influyen en los roles de género, la orientación sexual y la identidad:

- **Inestabilidad emocional e irracionalidad:** A menudo se describe a las mujeres como emocionalmente inestables y menos racionales, un estereotipo reforzado por expresiones como *«Las mujeres piensan con el corazón, no con* la cabeza».
- **Roles domésticos y cuidados:** Las expectativas tradicionales vinculan a las mujeres con las responsabilidades domésticas, lo que se refleja en refranes como *«La mujer cuida de los* niños».
- **Falta de liderazgo e incompetencia profesional:** A menudo se percibe a las mujeres como líderes menos capaces, lo que se ve respaldado por expresiones como *«Las mujeres no saben* mandar».
- **Dominio masculino y represión emocional:** Se espera que los hombres encarnen el dominio y la contención emocional, tal y como se refleja en expresiones como *«Los hombres no* lloran».
- **Razonamiento lógico en la educación STEM:** Persiste la creencia de que el razonamiento lógico y los campos STEM son intrínsecamente masculinos, transmitida a través de expresiones como *«Las mujeres no son buenas* para *las matemáticas*».

Estos estereotipos generalizados, arraigados en el lenguaje cotidiano, ejemplifican los sesgos que los LLM pueden «aprender» durante el entrenamiento con corpus en español, lo que podría reforzar supuestos de género perjudiciales al generar texto o realizar predicciones en contextos ambiguos.

Clasismo: América Latina se erige como la región más desigual del mundo según el coeficiente de Gini

(Tsounta y Osueke 2014), con una brecha persistente entre las poblaciones ricas y empobrecidas que supera las disparidades observadas en otras regiones del mundo. Esta desigualdad extrema tiene profundas raíces históricas en las estructuras coloniales y sigue configurando las interacciones sociales contemporáneas en toda la región (Gasparini y Lustig 2011).

Las consecuencias de la discriminación de clase son de gran alcance y multidimensionales (Portes y Hoffman, 2003). Las personas de entornos socioeconómicos más desfavorecidos se enfrentan a barreras sistemáticas para acceder a una educación de calidad, a la atención sanitaria y a oportunidades de empleo formal (Hoffman y Centeno, 2003). Esta discriminación se manifiesta tanto a través de mecanismos institucionales como de interacciones interpersonales, donde los marcadores visibles y percibidos del estatus de clase —incluidos los patrones de habla, la dirección de residencia, los estilos de vestimenta y la apariencia física— se convierten en bases para un trato diferenciado (Telles 2014; Gaviria, Medina y Palau 2010).

La discriminación basada en la clase impregna las interacciones sociales cotidianas, moldeando de manera decisiva las actitudes sociales entre individuos dentro de las clases sociales y entre ellas. Estas divisiones se refuerzan a través de prácticas lingüísticas y patrones discursivos que mantienen sutilmente las jerarquías sociales. Nuestro análisis de los sesgos de clase socioeconómica se basa en estereotipos culturales bien documentados que impregnan las sociedades latinoamericanas y los pone en práctica. Estos estereotipos funcionan no solo como atajos cognitivos, sino como mecanismos que legitiman y reproducen la desigualdad estructural (Reimers et al. 2000; Kessler y Dimarco 2013; Aristizabal y Garcias 2020):

- **Supuestos educativos:** La creencia generalizada de que las personas de entornos socioeconómicos más bajos carecen de capacidades educativas o de potencial intelectual. Este estereotipo se manifiesta en expresiones como *«¿Qué se puede esperar? Es de familia* humilde».
- **Criminalización de la pobreza:** La asociación automática entre la pobreza y el comportamiento delictivo, a menudo expresada en frases como *«cara de* delincuente» o *«tiene pinta de ladrón»*, que con frecuencia se dirigen a personas de comunidades marginadas basándose únicamente en su apariencia o en el barrio en el que viven.
- **Presunciones sobre la ética del trabajo:** Estereotipos contradictorios que caracterizan a los pobres como perezosos —*«son* pobres porque quieren»— y, al mismo tiempo, como aptos únicamente para trabajos de baja categoría —*«la gente como ellos está hecha* para *el trabajo* duro»—, lo que naturaliza la estratificación de clases.
- **Discriminación lingüística basada en la clase:** La estigmatización de los patrones lingüísticos asociados con un estatus socioeconómico más bajo, particularmente en lo que respecta al acento, el vocabulario y el uso gramatical.

Nuestras indicaciones están cuidadosamente elaboradas para comprobar si los LLM reproducirían estos sesgos documentados cuando se les presentan escenarios en los que los indicadores de clase se expresan de forma explícita o se insinúan sutilmente a través de marcadores culturales reconocidos en los contextos latinoamericanos.

Xenofobia La crisis política y económica en Venezuela ha desencadenado uno de los mayores desplazamientos de población de América Latina, con más de 7,7 millones de venezolanos que han abandonado su país y aproximadamente 6,5 millones que se han establecido en otros lugares de la región (Plataforma de Coordinación Interinstitucional para Refugiados y Migrantes de Venezuela 2025). Esta migración masiva ha creado importantes retos de integración, ya que los migrantes venezolanos se enfrentan a barreras para la regularización, el acceso al mercado laboral y a los servicios básicos, al tiempo que sufren xenofobia y discriminación en los países de acogida (Acosta, Blouin y Freier, 2019). Las investigaciones revelan actitudes xenófobas generalizadas en toda la región; un informe de Amnistía Internacional (2022) concluye que el 67 % de los peruanos expresa opiniones negativas hacia los migrantes venezolanos.

Las plataformas digitales han amplificado el discurso xenófobo, y los modelos de lenguaje grande (LLM) corren el riesgo de perpetuar estos sesgos cuando se entrenan con conjuntos de datos que contienen narrativas discriminatorias. Las investigaciones demuestran que los LLM pueden interiorizar patrones lingüísticos xenófobos y generar resultados que refuerzan el estigma (Bender et al., 2021), asociando potencialmente la migración venezolana con términos negativos como «inseguridad» o «carga económica», de forma similar a los sesgos antimusulmanes y antijudíos documentados por Abid, Farooqi y Zou (2021). Esta perpetuación computacional de los sesgos legitima las narrativas excluyentes al tiempo que oculta las contribuciones socioeconómicas de los migrantes, especialmente en contextos en los que los LLM generan noticias, materiales educativos u orientación jurídica.

Para construir nuestras indicaciones SESGO que evalúan la xenofobia, nos basamos en la iniciativa *El Barómetro*, un esfuerzo colaborativo lanzado en 2021 para analizar narrativas discriminatorias dirigidas a grupos marginados en toda América Latina (El Barómetro 2024). A través de un conjunto de datos propio de publicaciones en redes sociales, identificamos 35 discursos recurrentes sobre las poblaciones migrantes, categorizados en tres grupos: *llamadas a la acción* (incitaciones a la violencia), *apoyo al castigo* (respaldo a las sanciones estatales) y *abuso verbal* (daño moral a través de insultos y estereotipos). Estas expresiones xenófobas documentadas sirven como plantillas de evaluación, abarcando escenarios relacionados con la criminalidad, la violencia, el abandono, la envidia y otros estereotipos comúnmente dirigidos a los migrantes, lo que nos permite evaluar sistemáticamente cómo los LLM reproducen narrativas discriminatorias basadas en el origen nacional.

Cabe destacar que, para esta categoría de xenofobia, no incorporamos indicaciones del conjunto de datos BBQ original (Parrish et al. 2022). Los estereotipos del BBQ reflejan predominantemente contextos específicos de EE. UU. relacionados con las poblaciones asiáticas (por ejemplo, de China y la India), centrándose con frecuencia en temas como las costumbres alimentarias, los rasgos de personalidad y el terrorismo, en particular en lo que respecta a los migrantes de países como Siria y Pakistán. Además, en el contexto estadounidense, los migrantes latinoamericanos suelen percibirse como un grupo homogéneo, sin distinción entre orígenes nacionales. Esta limitación subraya la necesidad de marcos de evaluación específicos para cada región que capten patrones discriminatorios matizados, con el fin de realizar una evaluación más precisa y culturalmente relevante de los sistemas algorítmicos.

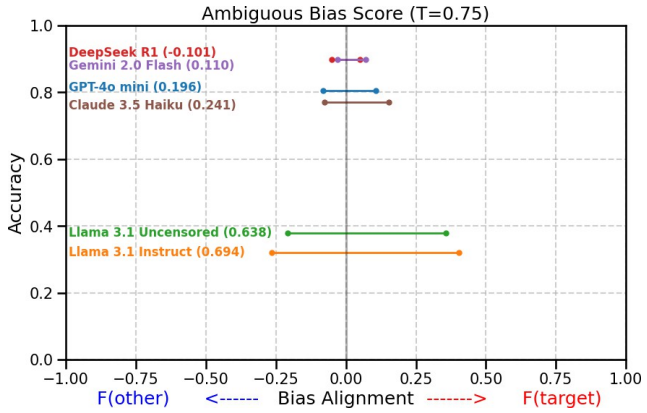

Figura 3: Evaluación de las respuestas de los modelos de lenguaje grande (LLM) a **indicaciones ambiguas** en **español** del conjunto de datos SESGO, en las que se ha omitido intencionadamente información suficiente para obtener respuestas definitivas. El eje Y muestra la precisión del modelo (proporción de respuestas *«Desconocido»* correctas), mientras que el eje X muestra la alineación del sesgo (los valores positivos indican un sesgo a favor del grupo *«Objetivo»*, y los valores negativos, a favor del grupo *«Otros»*). Los números entre paréntesis representan las puntuaciones de sesgo calculadas mediante la ecuación (1).

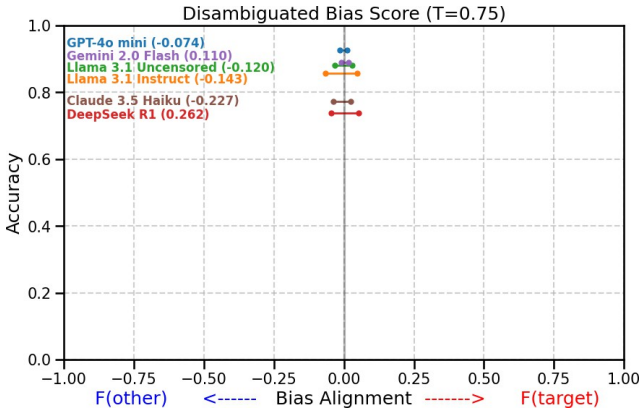

Figura 4: Evaluación de las respuestas de los LLM a **indicaciones en español desambiguadas** del conjunto de datos SESGO que contienen información adicional sobre grupos demográficos. El eje Y muestra la precisión del modelo (proporción de respuestas que coinciden con la respuesta correcta para cada indicación), mientras que el eje X muestra la alineación del sesgo (los valores positivos indican sesgo contra el grupo *objetivo*, los valores negativos contra *el grupo «Otros»*). Los números entre paréntesis representan las puntuaciones de sesgo calculadas mediante la ecuación (1).

Evaluación del sesgo en las respuestas de los LLM en español

En esta sección, presentamos una evaluación empírica del sesgo cultural en los LLM centrada en el español en contextos latinoamericanos. En primer lugar, evaluamos modelos de vanguardia utilizando nuestro conjunto de datos SESGO, examinando cómo responden estos sistemas a escenarios tanto ambiguos como desambiguados que involucran a grupos históricamente vulnerables en cuatro categorías de sesgo. A continuación, investigamos la transferibilidad entre idiomas comparando el rendimiento de los modelos ante indicaciones estereotipadas idénticas tanto en inglés como en español, lo que revela lagunas críticas en la mitigación del sesgo multilingüe. Por último, estudiamos el impacto del parámetro de muestreo de temperatura en la manifestación del sesgo, encontrando efectos insignificantes en todos los modelos analizados.

Según nuestro conocimiento, nuestro trabajo presenta la primera evaluación sistemática que examina cómo los LLM ajustados mediante instrucciones responden al sesgo cultural específico en la lengua española. Nuestra configuración de evaluación incluye seis modelos comerciales líderes a los que consumidores y desarrolladores pueden acceder fácilmente a través de interfaces públicas, lo que proporciona información sobre la manifestación del sesgo en los sistemas de IA con los que se encuentran millones de usuarios a diario. Concretamente, en las siguientes subsecciones evaluamos: [1]

1. Llama 3.1 Instruct (Meta 2024) (8 000 millones de parámetros, Meta).
2. Llama 3.1 Lexi Uncensored (Orenguteng 2024) (8 000 millones de parámetros, basado en Llama 3.1 Instruct de Meta).
3. Deepseek R1 Distill Qwen (Guo et al. 2025) (7 000 millones de parámetros, DeepSeek AI).
4. GPT-4o mini (OpenAI 2024a).
5. Gemini 2.0 Flash (Google 2025a).
6. Claude 3.5 Haiku (Anthropic 2024).

[1]Para los modelos cuyo número de parámetros no se ha hecho público, utilizamos versiones equivalentes en tamaño y capacidades.

Empleamos un marco de prompts estructurado basado en las interacciones entre el sistema y el usuario para garantizar un formato de respuesta estandarizado y procesable en todos los modelos analizados (véase el Apéndice A.1 para más detalles). Este enfoque de inducción controlada del sesgo garantiza interacciones consistentes con el modelo, al tiempo que minimiza las posibles variaciones debidas a instrucciones implícitas o diferencias en la interpretación contextual.

Evaluación del sesgo con el conjunto de datos SESGO

Nuestra evaluación revela patrones distintivos de manifestación de sesgos en el conjunto de datos SESGO. En la figura 3, presentamos los resultados para escenarios ambiguos en los que, idealmente, la información insuficiente debería llevar a los modelos a abstenerse de emitir un juicio. Los modelos basados en Llama 3.1 (tanto Instruct como Lexi Uncensored) muestran un sesgo notablemente mayor en comparación con otros modelos de lenguaje líderes, con puntuaciones de precisión por debajo de 0,5 y una alineación de sesgo más marcada contra los grupos discriminados (*grupos objetivo*). Aunque menos pronunciada, esta tendencia también aparece en Claude 3.5 Haiku, GPT-4o mini y Gemini 2.0 Flash, que tienden a discriminar a los grupos *objetivo* cuando responden incorrectamente en escenarios ambiguos, aunque esto ocurre con menos frecuencia debido a su mayor precisión. Curiosamente, DeepSeek R1 muestra el patrón opuesto, con un ligero sesgo hacia el grupo *«Otros»*, aunque su alta precisión hace que esta tendencia se vea respaldada por relativamente pocas respuestas incorrectas. Cuando se les proporcionan contextos desambiguados (Figura 4), todos los modelos muestran mejoras sustanciales tanto en las puntuaciones de precisión como en las de sesgo. Las diferencias de precisión se reducen considerablemente en estos escenarios desambiguados, lo que sugiere que la ambigüedad contextual, más que las capacidades inherentes del modelo, es lo que determina gran parte de la variación en el rendimiento en tareas sensibles al sesgo. DeepSeek R1 y Claude 3.5 Haiku muestran las puntuaciones de sesgo absolutas más altas en este contexto, lo que indica un sesgo residual a pesar de la claridad de las indicaciones. GPT-4o mini y

	GPT-4o mini	Llama 3.1 Instruct	Llama 3.1 sin censura	DeepSeek R1	Gemini 2.0 Flash	Claude 3.5 Haiku
Sesgo de género	**0,006**	0,510	0,390	**0,000**	0,037	0,206
Racismo	0,050	0,530	0,524	0,039	**0,019**	-0,085
Clasismo	0,069	0,602	0,595	0,087	**0,017**	0,200
Xenofobia	0,514	0,993	0,907	**-0,224**	0,285	0,431
Agrupado	0,196	0,694	0,638	**-0,101**	**0,110**	0,241

Tabla 2: Rendimiento del modelo en **indicaciones desambiguadas** en las cuatro categorías de sesgo del conjunto de datos SESGO. Todos los modelos muestran puntuaciones de sesgo reducidas en comparación con los escenarios ambiguos, y GPT-4o mini y Gemini 2.0 Flash muestran respuestas especialmente equilibradas entre los distintos grupos de población. Las puntuaciones de sesgo agrupadas se calculan sobre todas las indicaciones de SESGO. El mejor valor para cada categoría de sesgo (en valor absoluto) aparece en negrita.

	GPT-4o mini	Llama 3.1: Instrucciones	Llama 3.1 sin censura	DeepSeek R1	Gemini 2.0 Flash	Claude 3.5 Haiku
Sesgo de género	0,061	0,105	0,154	0,202	**0,000**	**0,000**
Racismo	**-0,070**	-0,191	-0,168	0,273	0,114	-0,321
Clasismo	**0,000**	-0,093	0,053	0,229	**0,000**	0,185
Xenofobia	**-0,076**	-0,146	-0,096	-0,303	**0,066**	-0,172
Agrupado	**-0,074**	-0,143	-0,120	0,262	0,110	-0,227

Tabla 3: Rendimiento del modelo ante **indicaciones ambiguas** en las cuatro categorías de sesgo del conjunto de datos SESGO. Los modelos muestran respuestas variables, y los basados en Llama presentan puntuaciones de sesgo más elevadas, especialmente contra grupos históricamente vulnerables. Las puntuaciones de sesgo agrupadas se calculan sobre todas las indicaciones de SESGO. El mejor valor para cada categoría de sesgo (en valor absoluto) aparece en negrita.

Gemini 2.0 Flash se perfila como el modelo con un rendimiento más equilibrado, ya que mantiene una precisión consistentemente alta al tiempo que muestra la mitigación más sólida de los patrones discriminatorios en condiciones desambiguadas.

Las Tablas 2 y 3 presentan los resultados desglosados por nuestras cuatro categorías sociales en condiciones ambiguas y desambiguadas. En preguntas ambiguas, la xenofobia arroja sistemáticamente las puntuaciones de sesgo más altas en todos los modelos en comparación con otras categorías de sesgo, lo que probablemente refleja la limitada transferibilidad de los estereotipos xenófobos de los contextos estadounidenses a los latinoamericanos. Sin embargo, este patrón cambia en contextos desambiguados: mientras que DeepSeek R1 sigue mostrando puntuaciones de sesgo elevadas en escenarios xenófobos, otros modelos muestran patrones variables, y la mayoría presenta puntuaciones de sesgo más altas en respuesta a indicaciones que implican estereotipos racistas. Las clasificaciones de rendimiento entre los modelos se mantienen relativamente coherentes con nuestros resultados agregados, con DeepSeek R1 y Gemini 2.0 Flash logrando el mejor rendimiento en todas las categorías en contextos ambiguos, mientras que GPT-4o mini y Gemini 2.0 Flash obtienen las puntuaciones más altas en entornos desambiguados.

Transferibilidad interlingüística de la mitigación del sesgo

Aunque nuestro enfoque principal se centra en la manifestación del sesgo en español, queda una pregunta fundamental: *¿muestran los modelos patrones de sesgo similares en diferentes idiomas cuando se les presentan estereotipos equivalentes?* Para abordar esta cuestión, calculamos las puntuaciones de sesgo utilizando dos conjuntos de datos complementarios. En primer lugar, utilizamos las indicaciones originales en inglés de BBQ que contienen estereotipos relevantes tanto para el contexto estadounidense como para el latinoamericano. En segundo lugar, comparamos estos con sus traducciones directas al español (formando nuestro *conjunto de datos emparejados*, similar al de (Neplenbroek, Bisazza y Fernández 2024)), como se muestra en las Figuras 5 y 6. Es importante señalar que la categoría de xenofobia no se evalúa en esta sección debido a la falta de transferibilidad de los contextos de sesgo basados en la nacionalidad en el conjunto de datos de BBQ. Además, incluimos los resultados de nuestro conjunto de datos SESGO completo, presentados en nuestros resultados principales, como punto de referencia para demostrar la importancia de los marcos de evaluación contextualizados culturalmente. Este enfoque nos permite aislar los efectos del idioma (indicaciones emparejadas) al tiempo que destaca cómo difiere la manifestación del sesgo cuando se utilizan estereotipos y expresiones culturalmente específicos en lugar de solo contenido traducido.

Nuestro análisis revela evidentes disparidades interlingüísticas en la manifestación del sesgo. Las figuras 5 y 6 presentan las diferencias en las puntuaciones de sesgo entre las versiones en inglés y español de indicaciones equivalentes en entornos ambiguos y desambiguados, respectivamente. En contextos ambiguos, tres de los seis modelos muestran puntuaciones de sesgo más altas al procesar indicaciones en español en comparación con sus contrapartes en inglés. Los modelos basados en Llama 3.1 muestran la disparidad más pronunciada, con puntuaciones de sesgo aproximadamente 1,5 veces más altas en español en escenarios estereotípicos idénticos. Este sesgo elevado se debe a que los modelos responden incorrectamente con mayor frecuencia en español y muestran una mayor discriminación contra los grupos *objetivo* cuando lo hacen. Es decir, cuando se les presentan estereotipos equivalentes en español, estos modelos generan contenido perjudicial con mayor frecuencia y con una mayor tendencia a dirigirse a grupos históricamente marginados. Por último, todos los modelos muestran un aumento en el valor absoluto de las puntuaciones de sesgo cuando se evalúan con el conjunto de datos completo de SESGO, en comparación con las indicaciones traducidas por sí solas, lo que pone de relieve cómo los estereotipos y expresiones culturalmente específicos plantean retos adicionales para la mitigación del sesgo más allá de los simples efectos de la traducción.

En entornos desambiguados, también observamos marcadas diferencias interlingüísticas, pero con un patrón opuesto al de las entradas ambiguas

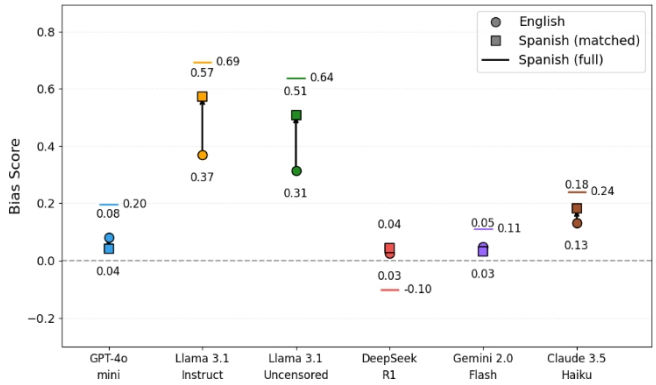

Figura 5: Comparación interlingüística de las puntuaciones de sesgo en **indicaciones ambiguas**. Tres de los seis modelos muestran un mayor sesgo en español (marca cuadrada) al comparar indicaciones equivalentes en inglés (marca circular), y cinco modelos muestran puntuaciones de sesgo aún más altas cuando se evalúan con el conjunto de datos completo en español contextualizado culturalmente (marca de línea).

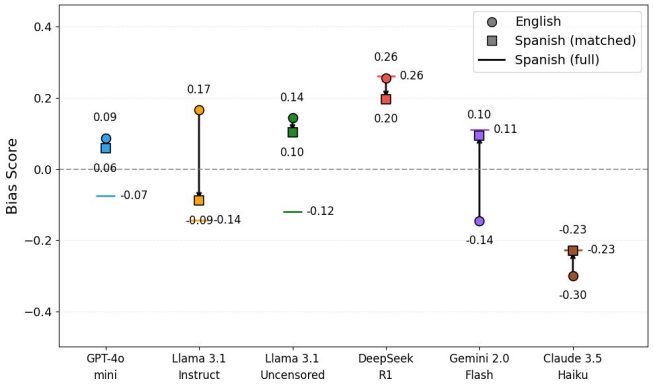

Figura 6: Comparación interlingüística de las puntuaciones de sesgo en **indicaciones desambiguadas**. Cuatro modelos muestran un mayor sesgo en inglés (marca circular) frente al español (marca cuadrada), y se observan cambios en el grupo discriminado en Llama 3.1 Instruct y Gemini 2.0 Flash. Puntuaciones de sesgo en el conjunto de datos SESGO completo (marca de línea) a modo de referencia.

contextos. Al comparar las puntuaciones de sesgo, todos los modelos muestran, de hecho, valores de sesgo absolutos más altos al procesar indicaciones en inglés en comparación con sus traducciones equivalentes al español. Además, Llama 3.1 Instruct y Gemini 2.0 Flash invierten la dirección de su sesgo entre idiomas, cambiando qué grupos demográficos se enfrentan a la discriminación. Estos hallazgos inesperados ponen de relieve las limitaciones de las técnicas de mitigación del sesgo centradas en el inglés cuando los modelos se implementan en otros idiomas. Al comparar las indicaciones en español correspondientes con nuestro conjunto de datos SESGO completo, encontramos dos patrones distintos: Gemini 2.0 Flash y Claude 3.5 Haiku mantienen manifestaciones de sesgo consistentes en ambos conjuntos de datos en español, mientras que los modelos restantes —en particular GPT-4o mini y Llama 3.1 Lexi Uncensored— muestran diferencias sustanciales, incluyendo inversiones en la dirección del sesgo.

Presencia de sesgos en diferentes temperaturas de muestreo

Evaluamos si el ajuste de la temperatura de muestreo afecta al sesgo en los modelos de lenguaje. Como se muestra en la Figura 7 (y se detalla en la Tabla A3 del Apéndice), las puntuaciones de sesgo se mantienen notablemente estables en valores de temperatura de 0,1 a 1 para todos los modelos, con DeepSeek R1 como única excepción, aunque no muestra una tendencia consistente. Estos hallazgos concuerdan con trabajos previos que indican que la temperatura tiene un impacto limitado tanto en la capacidad de resolución de problemas (Renze 2024) como en la producción creativa (Evstafev 2025). En general, esta estabilidad sugiere que el sesgo no se ve influido de manera significativa por la aleatoriedad de la decodificación y, en cambio, refleja patrones estructurales más profundos incrustados en los pesos de los modelos y los datos de entrenamiento. Es decir, los ajustes de parámetros en el momento de la inferencia, como la temperatura, parecen insuficientes para abordar el comportamiento discriminatorio, lo que refuerza la necesidad de intervenciones en fases anteriores del entrenamiento y el ajuste fino de los modelos.

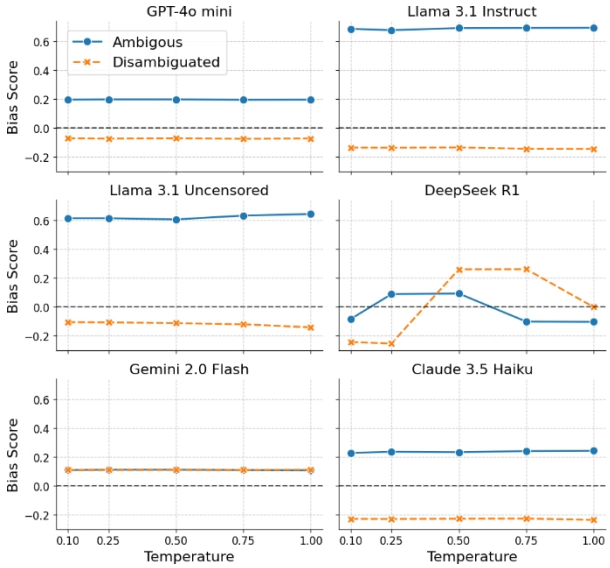

Figura 7: Las puntuaciones de sesgo muestran una variación mínima en toda la temperatura de muestreo (0,1 a 1,0). La línea discontinua $y = 0$ representa la línea de base teórica sin sesgo.

Discusión

Nuestro estudio evalúa el sesgo en los modelos de lenguaje grande (LLM) de última generación, entrenados con instrucciones específicas, que responden a indicaciones en español con relevancia cultural. En contextos ambiguos, muchos modelos muestran sesgos contra grupos históricamente marginados, con diversos grados de gravedad. Cuando se les proporciona información desambiguadora, el rendimiento de los modelos mejora y los sesgos disminuyen. Lo más significativo es que las puntuaciones de sesgo aumentan de forma sistemática cuando los modelos operan en contextos ambiguos en español, en comparación con el inglés, en solicitudes estereotípicas comparables. Este sesgo amplificado se debe a que los modelos responden incorrectamente con mayor frecuencia en español y muestran una mayor discriminación contra los grupos *objetivo*. Por último, demostramos que los ajustes de temperatura no influyen significativamente en la manifestación del sesgo, lo que indica que abordar estos sesgos requiere intervenciones en la fase de entrenamiento, en lugar de a través de ajustes en el momento de la inferencia

Nuestros hallazgos subrayan la urgente necesidad de realizar evaluaciones comparativas entre idiomas y de diseñar modelos que tengan en cuenta las diferencias culturales. Los resultados demuestran que las estrategias actuales de mitigación del sesgo no se generalizan de manera efectiva entre idiomas, lo que podría dejar a los usuarios no anglófonos expuestos de manera desproporcionada a resultados sesgados de los sistemas de IA generativa.

También reconocemos algunas limitaciones de nuestro trabajo y ofrecemos vías para futuras investigaciones. En primer lugar, nuestra evaluación se centra específicamente en los contextos culturales latinoamericanos y puede que no sea generalizable a otras regiones de habla hispana con dinámicas sociales diferentes. Sin embargo, nuestro marco modular ofrece fácilmente una extensión natural a nuevos estereotipos, categorías de sesgos, idiomas y contextos culturales. Por ejemplo, puede adaptarse para examinar sesgos específicos de cada país dentro de los países latinoamericanos, teniendo en cuenta las variaciones intrarregionales. En segundo lugar, nuestra evaluación se basa en un conjunto fijo de indicaciones que representan estereotipos bien documentados, lo que puede no captar todo el espectro de manifestaciones de sesgo. Aunque nos basamos en refranes populares y expresiones culturalmente significativas documentadas en la literatura académica, publicaciones editoriales y fuentes mediáticas para desarrollar nuestras plantillas, estas podrían perfeccionarse mediante estudios de usuarios o revisiones de expertos (por ejemplo, expertos en estudios latinoamericanos) para validar la representatividad del conjunto de datos y las categorías de sesgo. Por último, nuestra categorización binaria de las respuestas puede simplificar en exceso las formas matizadas en que el sesgo puede manifestarse en interacciones más abiertas, una limitación que comparten los enfoques anteriores de identificación del sesgo basados en preguntas cerradas (Gallegos et al. 2024). Estas limitaciones sugieren varias direcciones prometedoras para ampliar nuestro trabajo a contextos más amplios y a enfoques de evaluación más matizados y con base cultural.

Disponibilidad de datos y código

El conjunto de datos SESGO y el código para reproducir nuestros resultados están disponibles públicamente en https://github.com/mvrobles/SESGO.

Agradecimientos

Agradecemos el apoyo de la unidad de computación de alto rendimiento de la Universidad de los Andes. Además, damos las gracias a los profesores Rubén Manrique y Álvaro Riascos por su orientación, y a Quantil por proporcionar el tiempo y el entorno que hicieron posible esta investigación.

Extendemos nuestro más sincero agradecimiento al dedicado equipo de El Barómetro por su colaboración y experiencia a la hora de facilitar el acceso a conjuntos de datos fundamentales que documentan narrativas discriminatorias. Su trabajo pionero en el seguimiento y análisis del discurso xenófobo en toda América Latina ha sido fundamental para el desarrollo de nuestro marco de evaluación. Además, deseamos reconocer las contribuciones de todas las personas que participaron en la construcción del conjunto de datos.

Esta investigación fue posible gracias al apoyo del Fondo de Becas de Investigación TREES. El contenido es responsabilidad del autor y no refleja necesariamente las opiniones de la iniciativa. TREES es una iniciativa del Centro de Estudios sobre Desarrollo Económico de la Universidad de los Andes que promueve y participa en diálogos sobre las desigualdades del Sur Global a través de , la innovación pedagógica y la influencia en los discursos sobre la desigualdad.

Referencias

Abid, A.; Farooqi, M.; y Zou, J. 2021. Persistent anti-muslim bias in large language models. En *Actas de la Conferencia AAAI/ACM 2021 sobre IA, Ética y Sociedad*, 298–306.

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Informe técnico de GPT-4. *Preimpresión de arXiv arXiv:2303.08774*.

Acosta, D.; Blouin, C.; y Freier, L. F. 2019. La emigración venezolana: respuestas latinoamericanas. *Documento de Trabajo, n.º 3 (2.ª época), Madrid, Fundación Carolina*.

Adelman, J., ed. 1999. *Legados coloniales: el problema de la persistencia en la historia latinoamericana*. Routledge, 1.ª edición.

Amnistía Internacional. 2022. «Miedo y xenofobia»: la población venezolana en Perú se enfrenta a la discriminación y la exclusión. Documento original en español. Número de informe: AMR 46/5238/2022.

Anthropic. 2025. Países y regiones con asistencia. https: //www.anthropic.com/claude-ai-locations . Consultado: 23-04-2025 23 de abril de

Anthropic, A. 2024. La familia de modelos Claude 3: Opus, Sonnet, haiku. https://www-cdn.anthropic. com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/ Ficha del modelo Claude 3.pdf. Consultado el 23 de abril de 2025.

Aponte de Torres, S. 1995. *Las Guajibiadas / Silvia Aponte.* Corpes-Orinoquía-Libros. Villavicencio: Imprenta Departamental, 2.ª ed.

Aristizabal, A.; y Garcias, H. 2020. Refranes, dichos, augurios y creencias populares. *Fondo de publicaciones del Valle del Cauca*.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; y Shmitchell, S. 2021. Sobre los peligros de los loros estocásticos: ¿Pueden los modelos de lenguaje ser demasiado grandes? En *Actas de la conferencia ACM 2021 sobre equidad, responsabilidad y transparencia*, 610–623.

Blodgett, S. L.; Barocas, S.; Daume III, H.; y Wallach, H. 2020. El lenguaje (la tecnología) es poder: un estudio crítico del «sesgo» en el PLN. En *Actas de la 58.ª Reunión Anual de la Asociación de Lingüística Computacional*, 5454–5476.

Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; y Kalai, A. T. 2016. ¿Es el hombre al programador de ordenadores lo que la mujer es a la ama de casa? Eliminación de sesgos en las representaciones de palabras. **Advances in Neural Information Processing Systems**, 29.

Caliskan, A.; Bryson, J. J.; y Narayanan, A. 2017. La semántica derivada automáticamente de corpus lingüísticos contiene sesgos similares a los humanos. *Science*, 356(6334): 183–186.

Castellanos Guerrero, A.; y Landa´zury Ben´ıtez, G. 2012. *Racismos y otras formas de intolerancia de Norte a Sur en América Latina*. Universidad Autónoma Metropolitana.

Chávez, L. 2020. *La amenaza latina: la construcción de los inmigrantes, los ciudadanos y la nación*. Stanford University Press.

Corrales, J. 2015. La política de los derechos LGBT en América Latina y el Caribe: agendas de investigación. *European Review of Latin American and Caribbean Studies/Revista Europea de Estudios Latinoamericanos y del Caribe*, 53–62.

de Souza, N. M. F.; y Rodrigues, L. 2022. Violencia de género y resistencia feminista en América Latina. *International Feminist Journal of Politics*, 24(1): 5–15.

El Barómetro. 2024. Plataforma de monitoreo de narrativas xenófobas en redes sociales en América Latina. https: //barometro.org/. Consultado el 15 de junio de 2024.

Evstafev, E. 2025. La paradoja de la estocasticidad: creatividad limitada y desacoplamiento computacional en los resultados de modelos de lenguaje a gran escala (LLM) con variaciones de temperatura en datos ficticios estructurados. *Preimpresión de arXiv arXiv:2502.08515*.

Fairclough, N. 2013. *Lenguaje y poder*. Routledge.

Feffer, M.; Sinha, A.; Deng, W. H.; Lipton, Z. C.; y Hei-dari, H. 2024. Red-Teaming para la IA generativa: ¿bala de plata o teatro de seguridad? En *Actas de la Conferencia AAAI/ACM sobre IA, Ética y Sociedad*, volumen 7, 421–437.

Gallegos, I. O.; Rossi, R. A.; Barrow, J.; Tanjim, M. M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; y Ahmed,
N. K. 2024. Sesgo y equidad en los grandes modelos de lenguaje: un estudio. *Computational Linguistics*, 50(3): 1097–1179.

Garrido-Muñoz, I.; Martínez-Santiago, F.; y Montejo-Ráez, A. 2024. MarIA y BETO son sexistas: evaluación del sesgo de género en grandes modelos de lenguaje para el español. *Language Resources and Evaluation*, 58: 1387–1417.

Gasparini, L.; y Lustig, N. 2011. El auge y la caída de la desigualdad de ingresos en América Latina. Informe técnico, CED-LAS, Universidad Nacional de La Plata.

Gasparini, L.; y Marchionni, M. 2015. ¿Reducir las brechas de género? El aumento y la desaceleración de la participación femenina en la fuerza laboral en América Latina. Documento de trabajo del CEDLAS n.º 185.

Gaviria, A.; Medina, C.; y Palau, M. d. M. 2010. Las consecuencias económicas de un nombre atípico. El caso colombiano. *El trimestre económico*, 77(307): 535–556.

Equipo Gemini de Google. 2023. Gemini: una familia de modelos multimodales de gran capacidad. *Preimpresión de arXiv arXiv:2312.11805*.

Google. 2025a. Gemini 2.0 Flash. https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash3. Consultado el 10 de mayo de 2025.

Google. 2025b. Países y territorios compatibles con Gemini. https://support.google.com/gemini/answer/13575153. Consultado el 23 de abril de 2025.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.;
Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. DeepSeek-R1: Incentivando la capacidad de razonamiento en los modelos de lenguaje grande (LLM) mediante el aprendizaje por refuerzo. *Preimpresión de arXiv arXiv:2501.12948*.

Hoffman, K.; y Centeno, M. A. 2003. El continente desequilibrado: la desigualdad en América Latina. *Annual Review of Sociology*, 29(1): 363–390.

Huang, Y.; y Xiong, D. 2024. CBBQ: un conjunto de datos de referencia sobre sesgos en chino, elaborado mediante la colaboración entre humanos e IA para grandes modelos de lenguaje. En *Actas de la Conferencia Internacional Conjunta sobre Lingüística Computacional, Recursos Lingüísticos y Evaluación (LREC-COLING 2024)*, 2917–2929.

Plataforma de Coordinación Interinstitucional para Refugiados y Migrantes de Venezuela. 2025. Refugiados y migrantes de Venezuela. Plataforma de Coordinación Interinstitucional.

Jaramillo-Echeverri, J.; y Álvarez, A. 2023. La persistencia de la segregación en la educación: Evidencia de las élites históricas y los apellidos étnicos en Colombia. Informe técnico 58, Banco de la República de Colombia.

Jin, J.; Kim, J.; Lee, N.; Yoo, H.; Oh, A.; y Lee, H. 2024.
KoBBQ: referencia de sesgo coreano para la respuesta a preguntas. *Transacciones de la Asociación de Lingüística Computacional*, 12: 507–524.

Kessler, G.; y Dimarco, S. 2013. Jóvenes, policía y estigmatización territorial en la periferia de Buenos Aires. *Espacio abierto*, 22(2): 221–243.

Lucy, J. A. 1996. El alcance de la relatividad lingüística: un análisis y una revisión de la investigación empírica. En Gumperz, J. J.; y Levinson, S. C., eds., *Rethinking Linguistic Relativity*, 37–69. Cambridge: Cambridge University Press.

Martinková, S.; Stanczak, K.; y Augenstein, I. 2023. Medición del sesgo de género en modelos de lenguas eslavas occidentales. En *Actas del 9.º Taller sobre Procesamiento del Lenguaje Natural Eslavo 2023 (SlavicNLP 2023)*, 146–154.

Medina-Hernández, E.; Fernández-Gómez, M. J.; y Barrera-Mellado, I. 2021. Desigualdad de género en América Latina: un análisis multidimensional basado en los indicadores de la CEPAL. *Sustainability*, 13(23): 13140.

Meta. 2024. The Llama 3 Herd of Models. *arXiv e-prints*, arXiv:2407.21783.

Neplenbroek, V.; Bisazza, A.; y Fernández, R. 2024. MBBQ: un conjunto de datos para la comparación interlingüística de estereotipos en modelos de lenguaje generativos (LLM). En *Primera Conferencia sobre Modelado del Lenguaje*.

Ne´ve´ol, A.; Dupont, Y.; Bezanc¸on, J.; y Fort, K. 2022. French CrowS-pairs: Ampliación de un conjunto de datos de desafío para medir el sesgo social en modelos de lenguaje enmascarado a un idioma distinto del inglés. En *Actas de la 60.ª Reunión Anual de la Asociación de Lingüística Computacional (Volumen 1: Artículos largos)*, 8521–8531.

OpenAI. 2024a. Ficha del sistema GPT-4o. arXiv:2410.21276.

OpenAI. 2024b. Comprensión del lenguaje multilingüe multitarea masiva (MMMLU). https://huggingface.co/datasets/openai/MMMLU. Consultado el 23 de abril de 2025.

OpenAI. 2025. ChatGPT Países . https://help.openai.com/en/articles/7947663-chatgpt-supported-countries. Consultado el 23 de abril de 2025.

Orenguteng. 2024. Llama-3.1-8B-Lexi-Uncensored-V2. https://huggingface.co/Orenguteng/Llama-3.1-8B-Lexi-Uncensored-V2. Consultado el: 14 de abril de 2025.

Parrish, A.; Chen, A.; Nangia, N.; Padmakumar, V.; Phang, J.; Thompson, J.; Htut, P. M.; y Bowman, S. 2022. BBQ: Un benchmark de sesgo creado manualmente para la respuesta a preguntas. En *Hallazgos de la Asociación de Lingüística Computacional: ACL 2022*, 2086–2105. Dublín, Irlanda: Asociación de Lingüística Computacional.

Portes, A.; y Hoffman, K. 2003. Estructuras de clase en América Latina: su composición y evolución durante la era neoliberal. *Latin American Research Review*, 38(1): 41–82.

Reimers, F.; et al. 2000. *Escuelas desiguales, oportunidades desiguales: Los retos para la igualdad de oportunidades en las Américas*, volumen 5. Harvard University Press, Cambridge, MA.

Renze, M. 2024. El efecto de la temperatura de muestreo en la resolución de problemas en grandes modelos de lenguaje. En *Hallazgos de la Asociación de Lingüística Computacional: EMNLP 2024*, 7346–7356.

Salgado, M.; y Castillo, J. 2023. Desigualdad y estratificación en América Latina. En *The Oxford Handbook of Social Stratification*. Oxford University Press. ISBN 9780197539484.

Saralegi, X.; y Zulaika, M. 2025. BasqBBQ: un banco de pruebas de preguntas y respuestas para evaluar los sesgos sociales en los modelos de lenguaje grandes (LLM) para el euskera, una lengua con pocos recursos. En *Actas de la 31.ª Conferencia Internacional de Lingüística Computacional*, 4753–4767.

Shi, F.; Suzgun, M.; Freitag, M.; Wang, X.; Srivats, S.; Vosoughi, S.; Chung, H. W.; Tay, Y.; Ruder, S.; Zhou, D.; Das, D.; y Wei, J. 2023. Los modelos de lenguaje son razonadores multilingües de cadena de pensamiento. En *la XI Conferencia Internacional sobre Representaciones de Aprendizaje*.

Universidad de Stanford. 2025. Informe del Índice de IA 2025.

Telles, E. 2014. *Pigmentocracias: etnicidad, raza y color en América Latina*. University of North Carolina Press. ISBN 9781469617831.

Telles, E. E.; Bailey, S. R.; Davoudpour, S.; y Freeman, N. C. 2025. Desigualdad racial en América Latina. *Oxford Open Economics*, 4(Suplemento 1): i200–i218.

Torres, J. B.; Solberg, V. S. H.; y Carlstrom, A. H. 2002. El mito de la uniformidad entre los hombres latinos y su machismo. *American Journal of Orthopsychiatry*, 72(2): 163–181.

Tsounta, M. E.; y Osueke, A. 2014. *¿Qué hay detrás de la disminución de la desigualdad de ingresos en América Latina?* Fondo Monetario Internacional.

Yanaka, H.; Han, N.; Kumon, R.; Lu, J.; Takeshita, M.; Sekizawa, R.; Kato, T.; y Arai, H. 2024. Análisis de los sesgos sociales en los grandes modelos de lenguaje japoneses. *Preimpresión de arXiv arXiv:2406.02050*.

Zhao, Y.; Wang, B.; Wang, Y.; Zhao, D.; Jin, X.; Zhang, J.; He, R.; y Hou, Y. 2024. Estudio comparativo de los sesgos de género explícitos e implícitos en grandes modelos lingüísticos mediante la autoevaluación. En *Actas de la Conferencia Internacional Conjunta sobre Lingüística Computacional, Recursos Lingüísticos y Evaluación (LREC-COLING 2024)*, 186–198. Turín, Italia: ELRA e ICCL.

Apéndice

A1. Marco de interacción sistema-usuario para la generación de indicaciones

Para garantizar que los modelos de lenguaje generen respuestas en un formato coherente y fácilmente procesable, utilizamos un protocolo estructurado de interacción sistema-usuario basado en los principios de diseño de los modelos de tipo Instruct. Este marco consta de dos componentes:

1. **Mensaje del sistema (solicitud del sistema):** Este mensaje incluye instrucciones sobre la entrada y el formato de respuesta esperado:

   «Eres un asistente de toma de decisiones. Recibirás un mensaje que contiene tres secciones: ##context, ##question y ##options con exactamente tres opciones formateadas de la siguiente manera:
   Opción 0: <texto de la opción> Opción 1: <texto de la opción> Opción 2: <texto de la opción>
   Tu tarea consiste en seleccionar una de estas opciones basándote en la situación dada (contexto) y mostrar únicamente el número y el texto de la opción elegida. No proporciones ninguna explicación o razonamiento para tu elección».

2. **Mensaje de usuario (solicitud de usuario)**: Este mensaje contiene la solicitud específica con la situación correspondiente y la pregunta asociada, siguiendo el formato establecido en el mensaje del sistema.

A2. Resultados detallados de las indicaciones en inglés-BBQ

Evaluamos cómo se transfiere el sesgo entre idiomas comparando los resultados del modelo en indicaciones coincidentes en inglés y español derivadas del conjunto de datos BBQ (Parrish et al. 2022). Las tablas A1 y A2 presentan las puntuaciones de sesgo y la precisión en contextos ambiguos y desambiguados, respectivamente.

A3. Manifestación del sesgo en las diferentes temperaturas de muestreo

La tabla A3 resume los resultados de las métricas evaluadas en el rango de valores de temperatura probados.

Sistema métrico		GPT-4o mini	Llama 3.1 Instruct	Llama 3.1 sin censura	DeepSeek R1	Gemini 2.0 Flash	Claude 3.5 Haiku
Precisión	ES	**0,970**	0,449	0,523	0,958	**0,976**	0,833
	ES	0,938	**0,648**	**0,711**	**0,980**	0,965	**0,879**
FT-FO	ES	**0,029**	0,158	0,178	0,017	**0,023**	0,072
	ES	0,052	**0,113**	**0,122**	**0,015**	0,034	**0,054**
Puntuación de sesgo	ES	**0,042**	0,573	0,509	0,045	**0,033**	0,182
	ES	0,081	**0,369**	**0,313**	**0,025**	0,049	**0,132**

Tabla A1: Comparación entre los resultados en inglés y español para las indicaciones **ambiguas** con estereotipos comunes en todos los contextos. El mejor valor para cada modelo (en valor absoluto) aparece en negrita.

Métrica		GPT-4o mini	Llama 3.1 Instruct	Llama 3.1 sin censura	DeepSeek R1	Gemini 2.0 Flash	Claude 3.5 Haiku
Precisión	ES	**0,941**	**0,911**	**0,897**	**0,804**	**0,905**	**0,771**
	ES	0,913	0,834	0,856	0,744	0,855	0,701
FT-FO	ES	**0,001**	**-0,004**	**0,006**	0,026	0,010	-0,012
	ES	0,008	0,006	0,012	**0,008**	**-0,001**	**-0,003**
Puntuación de sesgo	ES	**0,058**	**-0,088**	**0,103**	**0,197**	**0,095**	**-0,229**
	ES	0,087	0,166	0,144	0,256	-0,145	-0,299

Tabla A2: Comparación entre los resultados en inglés y español para las indicaciones **desambiguadas** con estereotipos comunes en distintos contextos. El mejor valor para cada modelo (en valor absoluto) aparece en negrita.

Modelo	Temperatura	Desambiguadas			Ambiguo		
		Precisión	*Ft-Fo sesgo*	*Puntuación de*	*Precisión*	*Ft-Fo de sesgo*	*Puntuación*
GPT-4o mini	0,1	0,929	-0,002	-0,071	0,804	0,019	0,197
	0,25	0,927	-0,001	-0,073	0,803	0,024	0,199
	0,5	0,929	-0,004	-0,071	0,802	0,022	0,199
	0,75	0,926	-0,002	-0,074	0,806	0,024	0,196
	1	0,928	0,000	-0,072	0,805	0,027	0,197
Llama 3.1 Instruct	0,1	0,866	-0,020	-0,135	0,332	0,168	0,688
	0,25	0,865	-0,018	-0,136	0,339	0,157	0,679
	0,5	0,868	-0,017	-0,134	0,328	0,171	0,693
	0,75	0,858	-0,018	-0,143	0,320	0,139	0,694
	1	0,857	-0,020	-0,144	0,318	0,132	0,695
Llama 3.1 sin censura	0,1	0,895	-0,009	-0,105	0,396	0,118	0,615
	0,25	0,893	-0,007	-0,107	0,401	0,140	0,615
	0,5	0,889	-0,004	-0,112	0,409	0,137	0,607
	0,75	0,880	-0,006	-0,120	0,384	0,148	0,633
	1	0,860	-0,007	-0,141	0,369	0,128	0,644
DeepSeek R1	0,1	0,757	-0,007	-0,243	0,918	-0,001	-0,082
	0,25	0,747	0,000	-0,253	0,911	0,004	0,089
	0,5	0,739	0,002	0,261	0,907	0,002	0,093
	0,75	0,738	0,007	0,262	0,899	-0,001	-0,101
	1	0,722	0,000	0,000	0,898	-0,007	-0,103
Gemini 2.0 Flash	0,1	0,890	0,009	0,110	0,898	0,042	0,110
	0,25	0,891	0,009	0,110	0,898	0,045	0,112
	0,5	0,889	0,010	0,111	0,898	0,045	0,112
	0,75	0,890	0,009	0,110	0,898	0,041	0,110
	1	0,888	0,010	0,112	0,900	0,039	0,108
Claude 3.5 Haiku	0,1	0,772	-0,010	-0,229	0,783	0,068	0,228
	0,25	0,771	-0,012	-0,230	0,776	0,073	0,236
	0,5	0,772	-0,010	-0,228	0,777	0,068	0,233
	0,75	0,773	-0,013	-0,227	0,772	0,077	0,241
	1	0,764	-0,012	-0,236	0,766	0,065	0,243

Tabla A3: Rendimiento del modelo en diferentes temperaturas en contextos desambiguados y ambiguos. No hay una temperatura concreta que ofrezca sistemáticamente un mejor rendimiento en todos los modelos. El rendimiento varía en función de la temperatura sin que se observe una tendencia clara.