Sesgo cultural en grandes modelos de lenguaje: estado del arte y brechas de investigación para una IA más inclusiva

Hernán Javier Silva Sosa[1]
hjsilvasosa@uniciencia.edu.co

La inteligencia artificial (IA) se ha consolidado como una herramienta de uso generalizado en diversos sectores sociales, educativos y empresariales. No obstante, los sistemas contemporáneos de IA —y en particular los Grandes Modelos de Lenguaje (Large Language Models, LLM)— evidencian un sesgo cultural significativo, predominantemente orientado hacia valores, lenguas y contextos occidentales. Esta situación genera limitaciones en su aplicabilidad, representatividad y equidad frente a poblaciones culturalmente diversas.

El estudio propone realizar una revisión del estado del arte sobre el sesgo cultural en los LLM, con el propósito de identificar brechas en la investigación actual y proponer una agenda de desarrollo orientada a la construcción de modelos lingüísticos culturalmente inclusivos. Para ello, se efectuó una revisión en bases de datos académicas internacionales del periodo 2020 - 2025, caracterizado por un acelerado avance tecnológico en el ámbito de la IA generativa. Se establecieron criterios de inclusión que permitieron seleccionar investigaciones que evalúan la presencia, impacto o estrategias de mitigación del sesgo cultural. Tras un proceso de filtrado por título, resumen y palabras clave, se extrajeron datos relativos a las metodologías empleadas**,** los tipos de conjuntos de datos utilizados**,** los principales hallazgos**,** los impactos sociales identificados y las estrategias de mitigación propuestas.

El análisis permitió identificar patrones comunes**,** vacíos de conocimiento **y** limitaciones metodológicas en la investigación sobre LLM culturalmente inclusivos. Los resultados muestran que los LLM actuales se entrenan mayoritariamente con conjuntos de datos sesgados hacia contextos occidentales, lo cual se traduce en un bajo desempeño y una reducida representatividad en escenarios no occidentales. Asimismo, se evidenció la ausencia de marcos teóricos integrales que orienten el estudio del sesgo cultural, así como una limitada cantidad de investigaciones aplicadas en contextos culturalmente diversos. En consecuencia, se concluye que es necesario desarrollar marcos teóricos y metodológicos robustos que permitan comprender y mitigar el sesgo cultural en los LLM. La investigación futura deberá orientarse hacia la diversificación de los conjuntos de datos, la implementación de evaluaciones culturalmente sensibles y la promoción de colaboraciones interdisciplinarias e interculturales, con el fin de avanzar hacia la construcción de modelos lingüísticos más equitativos, inclusivos y culturalmente representativos.

Palabras Clave: Sesgo cultural, modelos, lenguaje inclusivo, inteligencia artificial

---

[1] Ingeniero Industrial, Mg. Administración de Empresas. Investigador Grupo Investigación GIII, CISE-UNICIENCIA

Cultural bias in large language models: State of the art and research gaps for more inclusive AI

Hernán Javier Silva Sosa[1]
hjsilvasosa@uniciencia.edu.co

Artificial intelligence (AI) has established itself as a widely used tool in various social, educational, and business sectors. However, contemporary AI systems—and particularly Large Language Models (LLMs)—evidence significant cultural bias, predominantly oriented toward Western values, languages, and contexts. This situation creates limitations in their applicability, representativeness, and fairness to culturally diverse populations.

This study proposes a systematic review of the state of the art on cultural bias in LLMs, with the aim of identifying gaps in current research and proposing a development agenda aimed at building culturally inclusive language models. To this end, a systematic review was conducted in international academic databases covering the period 2020–2025, a period characterized by accelerated technological advancement in generative AI. Inclusion criteria were established to select research that evaluates the presence, impact, or mitigation strategies of cultural bias in LLMs. After a filtering process by title, abstract, and keywords, data were extracted regarding the methodologies employed, the types of datasets used, the main findings, the identified social impacts, and the proposed mitigation strategies.

The analysis of the reviewed studies identified common patterns, knowledge gaps, and methodological limitations in research on culturally inclusive LLMs. The results show that current LLMs are mostly trained with datasets biased toward Western contexts, resulting in poor performance and limited representativeness in non-Western settings. Furthermore, a lack of comprehensive theoretical frameworks guiding the study of cultural bias was evident, as well as a limited amount of applied research in culturally diverse contexts.

Consequently, it is concluded that it is necessary to develop robust theoretical and methodological frameworks to understand and mitigate cultural bias in LLMs. Future research should focus on diversifying data sets, implementing culturally sensitive assessments, and promoting interdisciplinary and intercultural collaborations to advance the construction of more equitable, inclusive, and culturally representative language models.

Keywords: Cultural bias, models, inclusive language, artificial intelligence

[1] Ingeniero Industrial, Mg. Administración de Empresas. Investigador Grupo Investigación GIII, CISE-UNICIENCIA